Institut für Rechtsmedizin, Abteilung forensische Genetik, Charité-Universitätsmedizin, Berlin

D i s s e r t a t i o n

Establishment of a Y-chromosome specific extraction method for the separation of Y-chromosomal haplotypes from male DNA mixtures.

zur Erlangung des akademischen Grades

d o c t o r r e r u m n a t u r a l i u m

(Dr. rer. nat.)

im Fach Biologie - Genetik

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät I

der Humboldt-Universität zu Berlin

von

Diplom Biologin Jessica Rothe

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I

Prof. Stefan Hecht PhD

Gutachter/in:

1. PD Dr. Marion Nagy
2. Prof. Dr. Gottfried Fischer
3. Prof. Dr. Thomas Börner

Tag der Einreichung: 18.7.2013
Tag der mündlichen Prüfung: 23.04.2014

Abstract

Haplotype-specific extraction (HSE) allows the separation of diploid samples in their haploid components and was first demonstrated for the resolution of complex combinations of human leukocyte antigen (HLA) alleles. A similar problem is given with the analysis of mixed Y-chromosomal profiles which consist of two or more contributors. Also here an interpretation is only possible when the mixed profile can be separated in its single components of the corresponding contributors. At present, forensic communities as the International Society of Forensic Genetics (ISFG) and the German Stain Commission released guidelines and recommendations for the interpretation of DNA mixtures by statistical calculation. Nevertheless all solution statements refer to the interpretation of electropherograms which show an overlay of two or more profiles. Therefore the identification of the individual contributor still remains difficult and requires expertise. Here, HSE offers a new straight forward method to separate mixed profiles, consisting of haplotype markers like Y-chromosomal short tandem repeats (STRs). The advantage of the HSE approach in mixture analysis is the real physical separation of the individual DNA components before the amplification of the STR markers. In order to use the HSE technique for the separation of male DNA mixtures, it was first searched for Y-chromosomal single nucleotide polymorphisms (Y-SNPs) which were located near Y-STR markers. This search was carried out with current public human genome databases SNPs as well as databases for insertion-deletion mutations and polynucleotide polymorphisms. The outcome of this search showed that databases contain a substantial fraction of false entries. The main reasons for errors include sequencing or assembly errors, paralogous sequences, and private variants. In the course of the studies on the Y chromosome, a set of internal laboratory guidelines could be established in order to reliably identify false SNP entries in databases. With the validation of several SNPs as extraction sites, Y-chromosomal allele-specific extraction probes were designed which have been specific for one contributor of set up male mixtures. During extraction only complete matched probes are extended by a polymerase which results in the incorporation of biotinylated nucleotides. The synthesized and biotin labeled strand is separated by streptavidin coated magnetic beads. Finally, samples were analyzed by PCR coupled capillary electrophoresis for the detection of the extracted STR markers. First tests of a Y-chromosomal specific extraction of the STR loci DYS390, DYS437, DYS536 and DYS495 showed only little till no enrichment of the targeted alleles. Furthermore analyses with the AmpFLSTR® Yfiler multiplex PCR detected also non-targeted STR loci which are described in this study as HSE background signals. Test of different extraction protocol-steps and buffer components showed that these background signals are mainly caused by an unspecific biotinylation during HSE. A first improved separation effect of the Y-specific HSE could be obtained by the optimization of the used polymerase concentration. However the optimized

hybridization buffer did not lead to a complete elimination of the competing allele. Further optimization tests of different probe-parameters revealed the probe design as the key factor of successful HSE. The here presented results showed that new designed probes increase dramatically the separation success, often resulting in the extinction of the competitive allele. To validate these new results, HSE was extended to further six Y-chromosomal loci DYS389I+II, DYS437, DYS438, DYS439 and DYS635. In parallel, the thermodynamic characteristics as well as the individual hybridization-profiles of all tested probes have been analyzed with primer designing program Visual-OMP™. A comparison between the Visual-OMP™ simulated values and the results of the tested probe parameters showed that the extraction success of the new designed probes mainly depends of the relation of probe length and GC-contents. Because of the new gained knowledge about the influence of the probe-design on the separation success, probes for future HSE application can be developed faster and cost-effective. Furthermore the new prediction model for probe-specificity was also successful tested for the extraction of other genome-loci. In the last part of this work, a multiplex HSE approach was used to separate several STR markers simultaneously in one extraction reaction and therefore achieved the separation of one contributor Y-chromosomal haplotype.

Zusammenfassung

Die Haplotypspezifische Extraktion (HSE) ermöglicht die Trennung von diploiden Proben in ihre haploiden Komponenten und wurde erstmals entwickelt sowie angewandt für die Auflösung problematischer Allel-Kombinationen humaner Leukozytenantigene (HLA). Eine ähnliche Problematik ergibt sich auch für die Analyse Y-chromosomaler Mischprofile, bestehend aus der DNS zweier oder mehrerer Individuen. Auch hier ist grundsätzlich eine Auswertung nur dann möglich, wenn eine Trennung in die Einzelprofile (Haplotypen) der jeweiligen Verursacher erreicht werden kann. Verschiedene forensische Gesellschaften wie die deutsche Spurenkommission und die *International Society of Forensic Genetics* (ISFG) stellen speziell entwickelte mathematische Algorithmen sowie Empfehlungen für die Auswertung von Mischspuren zur Verfügung, die sich jedoch alle auf das bereits erstellte Mischprofil im Elektropherogramm beziehen. HSE bietet dafür einen neuen und direkteren Lösungsansatz, in dem die haploiden Y-chromosomalen DNS-Komponenten der einzelnen Individuen bereits vor der Analyse der individual spezifischen Marker separiert werden können und dadurch eine wirkliche physische Trennung erreicht wird. Für die Erstellung von individual spezifischen DNS-Profilen bedient sich die forensische Spurenanalyse der Vermehrung kurzer unterschiedlich langer Sequenzwiederholungen, den *short tandem repeats* (STRs). Um die Methodik der HSE für die Trennung von männlichen Mischspuren nutzen zu können, wurde zunächst nach Y-chromosomalen *single nucleotide polymorphisms* (SNPs) gesucht, die sich in der Nähe von Y-chromosomalen STRs, befinden (bis zu 80 kb Abstand). Für die Suche geeigneter Extraktionsorte wurden öffentliche humane Datenbanken genutzt, welche Informationen über SNPs aber auch Insertions- und Deletionsmutationen sowie Mehrfachnukleotid-Polymorphismen beinhalten. Eine Datenbankrecherche ergab, dass ein beachtlicher Teil aller Einträge fehlerhaft ist und zum Beispiel auf eine falsche Sequenzausrichtung oder auf paraloge Sequenzabschnitte zurück geführt werden können. Im Verlauf der Arbeit wurden laborinterne Richlinen zur sicheren und schnelleren Identifizierung falscher Datenbankeinträge erstellt. Nach der Validierung einiger SNPs als mögliche Extrationsstellen konnten allelspezifische Extraktionssonden erstellt werden, die nun gezielt nur die Marker der extrahierten DNS Komponente bzw. einer Person separieren sollen. Dabei werden im Hybridisationsschritt der HSE selektive nur komplett hybridisierte Sonden durch eine Polymerase verlängert. Gleichzeitig erfolgt während der Enzym-katalysierten Elongation eine Biotinylierung des neu entstehenden Stranges, welcher dann selektiv durch Streptavidin markierte magnetische Kügelchen extrahiert werden kann. Erste Durchführungen einer Y-chromosomal spezifischen Extraktion der Y-STR Marker DYS390, DYS437, DYS536 und DYS495 zeigten nur eine sehr schwache bis keine Anreicherung. Zusätzlich ließen sich auch Allele von nicht angereicherten STR Systeme

detektieren, welche in dieser Arbeit als Hintergrundallele bzw. -signale bezeichnet werden. Durch das Testen verschiedener Extraktionsprotokolle sowie Pufferkomponenten konnte die Entstehung solcher Hintergrundsignale auf eine unspezifische Biotinylierung während der HSE zurück geführt werden. Erste Verbesserungen der Anreicherung eines Alleles wurden durch die Optimierung der während der HSE eingesetzten Polymerasekonzentration erreicht. Allerdings ließ sich trotz des optimierten Hybridisierungspuffers das konkurrierende Allel nicht vollständig eliminieren. Im Folgenden wurden als Teil weiterer Optimierungsarbeit verschiede Sequenzvarianten einzelner Sonden getestet. Bei der Analyse verschiedener Sondenparameter an unterschiedlichen Loci des Y-Chromosoms wurde die Schlüsselstellung des Sondendesigns in der HSE-Technik deutlich. Die Ergebnisse zeigen, dass die neu entwickelten Sonden den Trennungserfolg der Mischprobe enorm verbessern und in einigen Fällen sogar zum Ausschluss des konkurrierenden Allels führen. Zur Validierung des neu entwickelten Probendesign wurden sechs weitere Extraktionsorte (DYS389I+II, DYS437, DYS438, DYS439 und DYS635) hinzugezogen, so dass insgesamt das Probendesign von 137 Sonden in mehr als 300 HSE Reaktion getestet werden konnte. Parallel dazu wurden mit Hilfe des Programms Visual-OMP™ die präzisen thermodynamischen Eigenschaften sowie das Hybridisationsprofil jeder Sonde ermittelt. Ein Vergleich der mit Visual-OMP™ simulierten Werte mit den Ergebnissen der getesteten Sonden ließ Rückschlüsse auf das allgemeine Bindungsverhalten der Sonden zu. Hierbei zeigte sich, dass der Extraktionserfolg der Sonde maßgeblich durch das Zusammenspiel von Sondenlänge und GC-Gehalt bestimmt wird. Durch dieses neu gewonnene Verständnis über den Einfluss der einzelnen Sondenparameter auf den Trennungserfolg der Mischprobe, können für künftige HSE Anwendungen Sonden effizienter erstellt und deren Wirksamkeit vorhergesagt werden. Zusätzlich konnte das neu entwickelte Vorhersage-Model der Sondenspezifität auch für weitere Extraktionsorte außerhalb des Y-Chromosoms bestätigt werden. Im letzten Teil der Arbeit konnte durch die Kombination verschiedener Sonden in einer Multiplex HSE die gleichzeitige Trennung mehrerer Y-chromosomaler Marker und damit die Trennung eines Haplotypen erzielt werden.

Table of contents

1 INTRODUCTION

1.1 Haplotype-specific extraction

Haplotyp-specific extraction is an allele-specific extension reaction (ASER) which allows the separation of diploid (or polyploid) DNA samples into their haploid components. HSE derives its power of discrimination from oligonucleotide primers (probes), which are designed with their 3'-termini specific to short variations, like single nucleotide polymorphisms (SNPs).

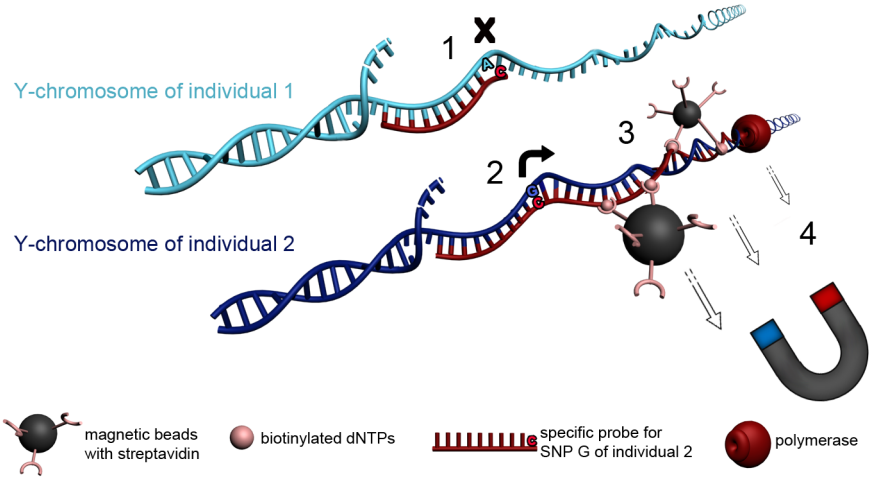


Figure 1-1: Schematic illustration of HSE for the separation of male DNA mixtures.
A male DNA mixture is shown, indicated by the presence of two different Y-chromosomes from individual 1 (light-blue) and individual 2 (dark-blue) which differ in a hypothetical SNP "A/G". In the example the Y-chromosomal fragment of individual 2 is isolated through the use of a selective probe that is designed to be matched (specific) to the targeted SNP allele "G". (1) The HSE-probe, shown in red, has a 3'-terminal mismatch ("C" vs. "A") for individual 1; therefore no extension occurs. (2) The probe matches the Y-chromosome of individual 2 completely ("C" vs. "G") and therefore will be extended. (3) During the extension reaction the polymerase incorporates biotinylated dNTPs. Streptavidin-coated magnetic beads are added to bind to the biotin-labeled DNA. (4) The DNA-biotin-streptavidin complex including the genomic template is captured through the use of a magnetic field.

The first step of HSE is the denaturation of the genomic DNA, followed by hybridization and enzymatic extension. During hybridization allele-specific probes anneal to the complementary sequence that is present on both of the two different haplotypes and form either a match or a mismatch with their 3'-terminus on the sequence variant. Then, thermostable *Taq* polymerase is used to initiate extension of the allele-specific probe that fully matches the underlying variant, including with its 3'-terminus. The polymerase therefore extends only completely matched probes and incorporates biotinylated nucleotides, resulting in the specific labeling of only one haplotype. Because *Taq* polymerase misses a 3' to 5'-exonuclease activity, any 3'-terminal mismatches between probe and haplotype are not enzymatically removed and therefore cannot be extended [1]. In a final

step the biotinylated strand of the targeted allele as well as its underlying genomic DNA template are separated by an extraction step with streptavidin-coated magnetic beads (Figure 1-1). Therefore, HSE allows the separation of large and contiguous chromosomal fragments or haplotypes.

With the haplotype-specific extraction (HSE), a patented technology from Dapprich and Cleary [2], individual alleles or haplotypes can be collected from mixed genomic DNA samples. HSE applications were first described by Nagy *et al.* 2007 [3], regarding the resolution of two ambiguous human leukocyte genotypes: HLA-B*47, *56 and HLA-B*44, *56. In this work, the four separated haplotypes were unambiguously sequenced and two novel alleles, HLA-B*4440 and HLA-B*5613, were identified. To date, numerous ambiguous HLA and killer-cell immunoglobulin-like receptors alleles have been resolved by HSE [4-9]. However, HSE offers also additional applications for different scientific approaches. Gabriel *et al.* 2006 [10] used the HSE technique for a transposon-specific DNA extraction in *Saccharomyces cerevisiae*. Hereby transposon-associated sequences were isolated and analyzed by a microarray-based comparative hybridization assay to map the repertoire of endogenous transposon locations in different yeast strains. One further interesting application for HSE is the region specific extraction (RSE) and was used by Gupta and colleagues to identify a causative mutation that leads to a developmental defect. They used the technique to enrich large (>300 kb) regions of the zebrafish genome for subsequent next generation sequencing analysis on the Illumina platform [11]. Further validation of RSE for next generation sequencing platforms would be very advantageous, since other cannot enrich large chromosomal fragments bridge across repetitive regions or regions that include large stretches of unknown sequence, and are not yet cost effective.

Allele-specific extension reactions (ASERs), or allele-specific primer extensions, are used in a broad range of applications. They are often the assay of choice for the detection of single nucleotide polymorphisms (SNPs) for medical usage, in disease association studies, and in routine molecular diagnostics [12]. Depending on the application, several ASER techniques have been developed over the years, which comprise different detection assays like autoradiography, bioluminescence with photoproteins, dye-terminator incorporation assay (cycle sequencing), luciferin-based chemiluminometric assay (Pyrosequencing), high resolution melting (HRM) analysis, color complementation, and others [13-18].

One derived form of ASER is the use of allele-specific primers in an allele-specific amplification (ASA) or allele-specific PCR (ASP), which has been first demonstrated by Newton *et al.* 1989 [19] and Wu *et al*. 1989 [20]. Here only the wildtype or the mutant allele is selectively amplified through the use of

3'-specific primers. Reactions with mismatched primer show reduced amplification efficiency and no PCR product should be detected. An ASA method developed by Chehab *et al.* 1989 [15] used different fluorescent dyes for mutant and wildtype extension primers, thus avoiding the separation of ASA reaction and electrophoresis.

In the second basic form of ASER, only one allele-specific primer or probe is extended rather than amplified. This is commonly called allele-specific primer extension (ASPE) or also simply primer extension (PEX or PEXT). Hereby PEX are often used for single base extension reactions (SBE) in which a primer is designed to hybridize to the target sequence immediately 5' of the polymorphic site. Early SBE reactions have been described by Sokolov *et al.* 1990 [18] and Kuppuswamy *et al.* 1991 [21], who set up PEX with only one of the four with radioisotopes labeled dNTPs [18, 21].

Today's advanced SBE reactions generally use fluorescently labeled dNTPs, which can be set up together in one reaction and are often available as standard reaction kits, e.g. the SNaPshot assay [22]. One other interesting assay was developed by Kornher and Livak [23], which uses the insertion of mobility-shifting analogs during the PEX. In their assay one of the four natural dNTPs is replaced with a mobility-shifting analog (MSA) and the extension primer is elongated for only a few dNTPs to a defined length. Thus, when DNA-fragments are compared by electrophoresis, strands that are of the same length, but differ in the number of analog residues, will migrate at different rates.

Alternatively to exploiting the primer position adjacent to the SNP – as done in SBE reactions, an allele-specific primer can be used whose 3-end is complementary to the nucleotide at polymorphic site. The result is, similar to ASA, only the primers whose 3'-end perfectly match the interrogated sequence are extended by the polymerase. The HSE method uses allele-specific probes but forgoes DNA amplification – it therefore corresponds most closely to a PEX reaction with a 3-end specific primer. However, also PEX with allele-specific primers show manifold assay techniques (Table 1-1). For example, Konstantou *et al.* 2007 [24] describe an assay that is similar to HSE in which the extension product is labeled with biotin-dUTPs or digoxigenin-UTPs and then detected by a bio-chemiluminometric hybridization assay. Also Papanikos *et al*. 2012 [17] use biotinylated extension products, but they describe a visual detection method which is more suitable for clinical diagnostics. Here successful primer extension is detected by the binding of the biotin-labeled extension product to gold nano particles, which form a visible accumulation when the reaction product is passed through a diagnostic membrane. In contrary to HSE, the PEX assays from Konstantou and Papanikos are set up with amplified PCR products and the extension reaction is carried out with several cycles similar to a PCR. Another PEX assay more analogous to HSE is described by Chan *et al.* 2011 [25]. This group sets up the primer extension reaction for only 30 min. at 58°C without further amplification steps. But also here PEX starts from a PCR product and uses fluorescently labeled dyes for extension

and detection. The PEX assay developed by Cai *et al*. 2010 [14] combines common PCR and primer extension in one reaction and uses LCGreen-mediated HRM analysis for the SNP detection.

Table 1-1: Different applications of allele-specific extension reaction.

publication	Kornher and Livak, 1989 [23]	Kuppuswamy *et al.*, 1991 [21]	Konstantou *et al.*, 2007 [24]	Cai *et al.*, 2010 [14]	Chan *et al.*, 2011 [25]	Papanikos *et al.*, 2012 [17]
assay abbreviation	**MSA**	**SBE**	**PEX**	**PEX**	**PEX**	**PEX**
principle of the method	SNP is detected by altered electrophoretic mobility, caused by the insertion of different nucleotide analogs	uses four identical reactions, each with one of the four dNTPs - only one shows an extensions depending on SNP	only perfect matches are extended and the new strand is labeled with biotin	SNP detection by HRM, which combines in one reaction amplification of normal PCR product and an extension product, only perfect matched primers are extended	only perfect matches are extended and the new strand is labeled with fluorescence marker	only perfect matches are extended, biotin is incoperated in the new strand and then in a further assay detected
target	PCR product	PCR product	PCR product	20ng genomic DNA	PCR product	PCR product
characteristics of the extension primer	50nM, 20nt, end labeled with γ-^{32}P-ATP	1µM	0.25µM, >19nt plus adapter	5µM excess forward primer, 25nM reverse primer, 5µM extension primer, <14nt	primer spotted on coated glass slides	0.25µM, >20nt plus adapter
position of primer to SNP	primer is placed before SNP	primer is just before SNP	3'end of primer at SNP	3'end of primer at SNP	3'end of primer at SNP	3'end of primer at SNP
polymerase	0.4U/µl Taq polymerase	2U Taq polymerase	0.25 U Vent (exo-) DNA polymerase	kit Premix Ex Taq master mix and LCGreen Plus for HRM	0.75x Titanium™ Taq polymerase	1U vent (exo-) DNA polymerase
dNTPs	10µM dTTP or biotin-11-dUTP 10µM dCTP or 5-(bio-ACAP3)dCTP 10µM dATP 10µM dGTP	10µCi/µl (α32-P)dNTPs, only one per reaction	1.25µM dTTP and 1.25µM biotin-dUTP 2.5µM dATP 2.5µM dATP 2.5µM dGTP		0.25µM Cy5-dCTP 0.25µM dTTP 0.25µM dATP 0.25µM dGTP 0.125µM ddNTPs	5µM dTTP or biotin-11-dUTP 10µM dCTP 10µM dATP 10µM dGTP
buffer	50mM Tris-HCL pH9, 5mM $MgCl_2$	10mM Tris-HCl (pH 8.3), 50mM KCl, 5mM $MgCl_2$, 0.001% gelatin	20mM Tris-HCl, 10mM $(NH_4)_2SO_4$, 10mM KCl, 0.1% Triton X-100, (pH8.8), 2mM $MgCl_2$		10mM Tris-HCl, 50mM KCl, 1mM $MgCl_2$, 4% DMSO	20mM Tris-HCl (pH8.8), 10mM $(NH_4)_2SO_4$, 10mM KCl, 0.1% Triton X-100, 4mM $MgSO_4$
reaction parameters	70°C for 5 min	one cycle of 95°C for 2min, 60°C for 2min, 72°C for 2min	3 cycle of 95°C for 10sec, 65°C for 10 sec, 72°C for 30 sec	first amplification: 50 cycles of 95°C for 6 sec, 64°C for 12 sec. extension reaction with 30°C for 2min, 40°C for 2min and 60°C for 1min	58°C for 30min	10 cycle of 95°C for 15 sec, 60°C for 15 sec, 72°c for 15 sec
detection method	autoradiography	autoradiography	(bio)chemilumino-metric hybridization assay	HRM	detection of fluorescence	dipstick method

Green indicates reaction steps similar to HSE, blue reaction steps different to HSE.

However, for all applications of ASA and PEX, the reliability and specificity of the approach is dependent on the ability of the *Taq* polymerase to discriminate between the extension of mispaired and canonically paired primers. Based on the crystal structure of the *Taq* polymerase, it belongs to the well studied polymerase family A, which includes for example also the polymerase I and the Klenow fragment (KF) of *Escherichia coli.* In general, their global architectures resemble the shape of a "right hand", with three distinct subdomains forming the "palm", "thumb" and "fingers". The function of the "palm" domain appears to be the catalysis of the phosphoryl transfer reaction by a two-magnesium-ion catalyzed mechanism. The "fingers" domain includes important interactions with the incoming nucleoside triphosphate as well as the template base, whereas the "thumb" plays a role in positioning the duplex DNA [26, 27]. Furthermore the *Taq* polymerase contains also a 5-

nuclease and an inactive 3'-5' exonuclease domain. The KF misses the 5'-nuclease domain but shows 3'-5' proofreading activity. An analysis of the proofreading mechanism for the KF indicates a competition between the exonuclease and the polymerase active site for the 3'-end of the primer and a rapid shuttling of the primer terminus between these two sites. The 3'-exonuclease active site binds single stranded DNA, whereas the polymerase active site binds duplex DNA (i.e. correctly base-paired duplex DNA). Carver etal. 1994 [28] demonstrated that GC-rich sequences bound stronger to the polymerase domain whereas AT-rich termini showed greater partitioning in the exonuclease side and an increased rate of exonucleolysis. Mismatched base pairs however destabilize the duplex DNA and thereby enhance the binding of the 3-single-stranded DNA to the exonuclease active site, which normally would then excise any incorrectly matched bases [27, 28]. Even though the 3'-5' exonuclease catalytic site of the *Taq* polymerase has been destroyed and the size of its domain reduced, the contact with the polymerase domain as well as the distance between the polymerase domain and 3'-5' exonuclease domain remains similar to the homologous KF. Therefore it is supposed that the *Taq* polymerase uses a similar mechanism (i.e. shuttling between polymerization and proofreading mode) to successfully discriminate mispaired versus correctly paired bases [26, 29]. Furthermore, studies of different buffer conditions for several polymerases showed that reaction conditions, such as pH, concentrations of dNTPs and magnesium ions, can greatly affect the fidelity of the DNA polymerase [30]. However, because *Taq* polymerase binds with similar affinities to matched and mismatched primer templates, the non-specific extension of a 3'-nucleotide mismatch is often a significant fraction of the overall extension product which can be controlled with optimized probe design [31, 32]. Therefore, the thermodynamic considerations of optimal allele-specific probe design must be considered. Early studies already showed that internal mismatches and mismatches near the 5'-end have no significant effect on ASA specificity, but should ideally be present in the first four positions from the primer 3'-end [33-35].

The studies of Onodera *et al.* 2007 [36] give an overview about the 3'-end triplet frequencies in primers for optimally specific PCR. They recommend that primers should be designed with SWS, WSS or TTS as 3-end triplets, whereas the triplet types, WWW, CGW, and GGG (with S stands for C or G and W stands for A or T) should be avoided. In contrast to Onodera an early evaluation of Newton *et al.* 1989 [19] correlates ASA success with the mismatch type and suggests that successfully allele-specific primer discrimination occurs considerably more reliably through 3' purine-purine or pyrimidine-pyrimidine mismatches, as opposed to 3' purine-pyrimidine (i.e. G:T or C:A). Ayyadevare *et al.* 2000 [37] reported that template discrimination was optimal when the nucleotides T, G or C but not A occupied the 3'-end position of a primer. This group further went on to investigate the effect on successful template discrimination that is generated by the nucleotide identity at the

penultimate (-1, next to the last) position relative to the 3'-end of the primer and reported that amplification efficiency was reduced, when T or A occupied this position. Furthermore, they observed a reduced discrimination of 3'-end G:T mispairing, which was also reported by other groups and consists with the observation, that G:T base pairing can contribute to stable duplex formations [33, 38, 39]. However, several studies about optimizing 3'-end primer design and selecting the most efficiently discriminating types of mismatches also reveal discrepancies between their respective findings. For example, Bottema and Sommer [40] observed successful mismatch discrimination for all tested combinations, A:C, G:T, C:A and T:G, whereas Kwok *et al.* 1990 [33] found differential amplification only for the mismatches A:G, G:A, C:C and A:A. Obviously, discrepancies among different studies may have been caused by different experimental protocols but they typically are reported as being primarily due to primer-sequence-dependent differences [37, 41]. Successful allele-specific primer extension depends on the ability of the polymerase to recognize and extend a correctly matched duplex over a duplex containing a mismatch. This ability to discriminate between alleles is influenced by the duplex stability of mismatches and their location, by the reaction kinetics of association and dissociation of primer-template duplexes at the annealing and extension temperatures, by the buffer conditions, and last but not least by the type of polymerase used and its evolutionary origin and purpose (additional effects that can be generated through the use of nucleotide analogs or modified primers, such as locked nucleic acids (LNA) are not discussed here). The duplex stability of primer-template is determined primarily by the complementary bases that make up the duplex and by its GC content, but also by sequence dependent interactions between target and probes which is described as base stacking [42-44]. Nearest-neighbor stability parameters have been introduced to account for sequence-dependent effects in DNA stability, and statistical thermodynamic and kinetic calculations have been developed to simulate DNA multiplex hybridization reactions under multi-state coupled equilibrium conditions. The thermodynamic effects of primer (or probe) nucleotides and the effects that their "nearest neighbors" on the overall stability of base-pairing were quantified, amongst others, by SantaLucia *et al.* [45], who determined the enthalpic and entropic contributions of nearest neighbor match and mismatch sequence composition, base stacking, and mismatch geometry [38, 46-48]. These nearest neighbor parameters are considered in most oligonucleotide design and simulation programs (e.g. Oligo Analyzer, Oligo Explorer or Primer 3), including the software platform used extensively in this work, Visual OMP™ (DNA Software, Inc., Ann Arbor, MI, USA).

1.2 Forensic DNA analysis

One major task of forensic science is to trace individual-specific markers from biological evidences in order to identify single persons and to in- or exclude suspects from crime scene. The first genetic evidence used in court came from the ABO blood groups, discovered by Karl Landsteiner in 1901. The use of DNA for forensic and human identification purposes began with the work of Alec Jeffreys in the early 1980s. He showed that VNTR (variable number of tandem repeats) minisatellites could be used to track genetic inheritance and differentiate between people [49]. This early DNA testing was performed by the technique of restriction fragment length polymorphism (RFLP) analysis and used highly variable minisatellites with typical repeat units of 8-100 bp. For RFLP analysis, genomic DNA is digested into smaller fragments (< 12 kb) with one restriction enzyme (e.g. *Hae* III). Then the fragments, which contain variable numbers of repeats, are detected by southern blots and for example by radioactively or digoxigenin labeled VNTR probes. The patterns that are produced look like barcodes and were termed DNA fingerprints because they are so variable among the individuals like the friction ridges of the human finger. Although the RFLP technique allows for highly reliable discrimination among different individuals (~1 in millions or greater with four loci), it is time consuming and not suitable for degraded or low copy number (LCN) DNA samples.

A second revolution of DNA testing started after the discovery of microsatellites and the development of the polymerase chain reaction [50]. Microsatellites consist of repeat units in the range of only 1 to 7 bases and allow therefore the amplification of smaller fragment and the analysis of degraded and low copy number (LCN) DNA samples. Because microsatellites have been proven to be more practical for laboratory routine, they are now commonly used and called short tandem repeats (STRs). STR markers between individuals differ in general by the number of repeats. Typical STR mutation rates are between 10^{-3} – 10^{-4} per locus per generation [51]. The major mechanism of mutation for microsatellites occurs as a result of DNA replication slippage which displays an out-of-register realignment after the dissociation of the replicating DNA strands. The new wrongly aligned strands result in either a gain or a loss of generally one repeat unit, depending on whether a loop is formed on the nascent or the template strand, respectively [52, 53].

The developments of fluorescent dyes and high resolution systems of fragment lengths by capillary electrophoresis (CE) allow a simultaneous detection of several STR loci in a rapid, exact and automated process. At the present time, forensic laboratories use commercially available multiplex kits for the determination of a selected set of STR markers, and technology is continually being improved by increasing the number of simultaneously amplified marker. For example, the most recent product from Life Technologies / Applied Biosystems is the AmpFLSTR® NGM SElect™ kit,

which combines the detection of 16 loci and the amelogenin locus for sex determination [54, 55]. Electrophoresis of the multiplex STR PCR results in an electropherogram with size-separated PCR products shown as color-separated panels (Figure 1-2).

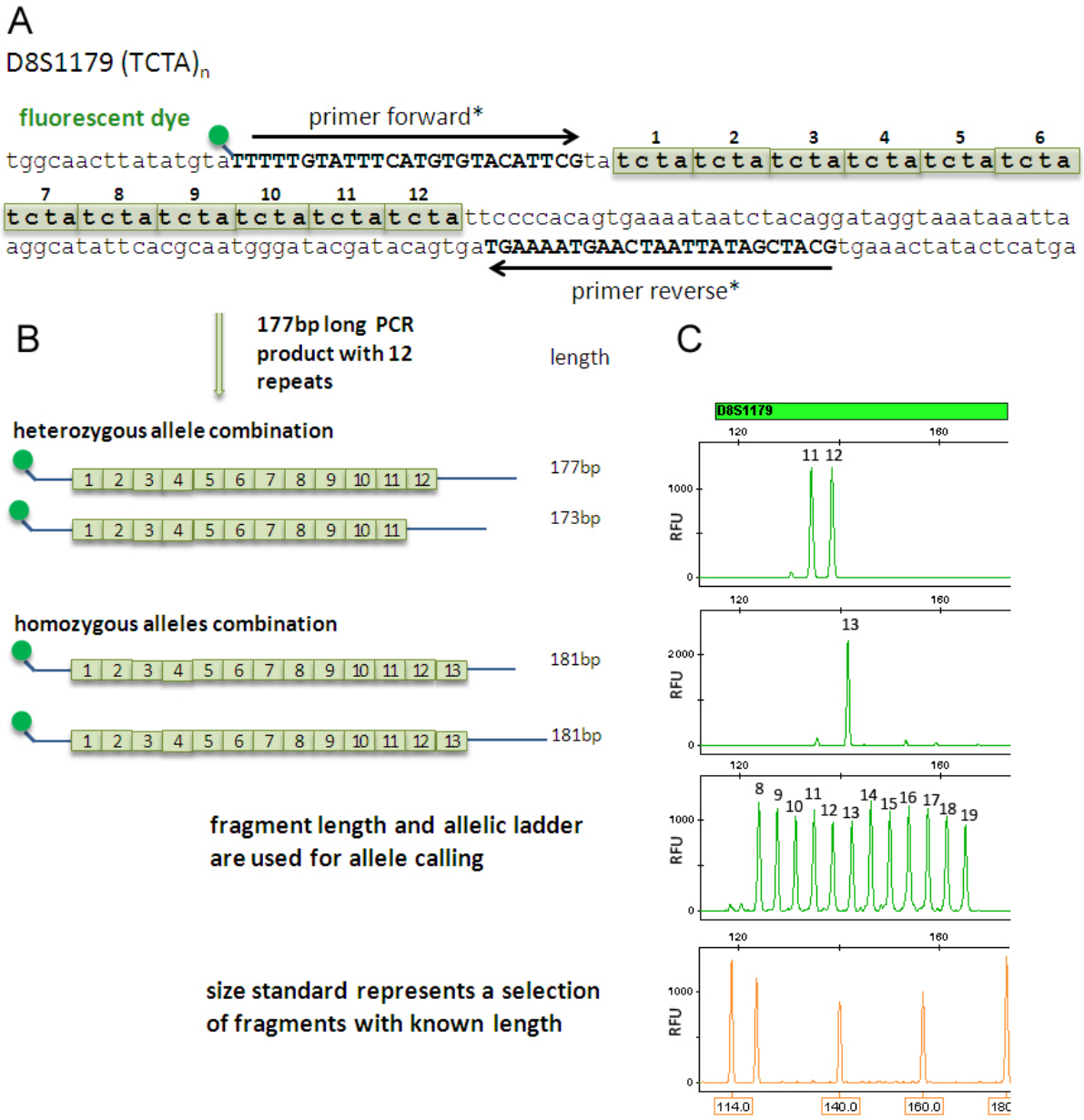


Figure 1-2: Principle of STR typing.
(A) PCR primers (arrows) anneal to sequences up- and downstream of the variable STR repeat region. One of the PCR primers is labeled with a fluorescent dye. (*) Primer sequences are taken from Barber and Parkin 1996 [56] (B) The variable number of repeats within the STR maker generates PCR products, of different length for different DNA samples. (C) Labeled PCR products with different sizes appear as peaks with different positions in the electropherogram. The intensity of the emitted light is indicated in relative fluorescence units (RFU). The peaks are sized by using an internal standard of defined length and correlated to one specific allele with a precise number of repeats. The figure is redrawn from Butler, 2009 [50].

Hereby each STR allele is presented as a distinct peak with a precise fragment length which can be calculated by the use of an internal standard of defined length. The genotype, or the characteristic

number of repeats for the DNA sample of a specific individual, can then be designated by a comparison of the peak sizes obtained with “size bins” for each known allele, which are confirmed in each run by an allelic ladder (Figure 1-2, chapter 2.12). The use of different dyes allows the use of multiple bin sizes and therefore the multiplexing of a high number of markers. This new technology has turned STR profiling into a common tool in forensic investigations and initiated the rapid growth of national DNA databanks and population databases.

Population databases have become essential tools for the calculation of allele frequencies and thus of probabilities for random matches to occur between different DNA samples. A DNA database is a collection of computer files, containing entries of DNA profiles that can be searched to look for potential matches. They have demonstrated their ability to serve as valuable tools aiding law enforcements investigations, for example by making association between groups of unsolved cases. In Germany the *DNA Analysedatei* (DAD) database was founded in April 1998 with five core STR systems, which were extended in 2001 to eight markers. In 2009, these DAD markers have been extended by a council resolution of the European Union by five further ESS markers (European Standard Set) [57]. In the case of a DNA match between suspect and evidence, a random match probability can be calculated by multiplying all single allele frequencies. In doing so, the use of the product rule assumes that all tested STR loci are independent and in full linkage equilibrium, i.e. in random association. The match probability describes the likelihood that a second, unrelated person exists with exactly the same DNA profile. For example, match probabilities calculated from 16 STR loci are about 1×10^{-23} and therefore easily provide convincing specificity for individuals on a world-wide range[1].

The establishment of databases and their use for forensic investigation on national and international levels require agreement about a core set of commonly used markers as well as a permanent quality assurance and control. Therefore, around the world, many organizations work on laboratory guidelines, quality standards and accreditation programs. For example the Scientific Working Group on DNA Analysis Methods (SWGDAM) has published several guidelines for STR typing which are available on their homepage. In 1995, the European Network of Forensic Science Institutes (ENFSI) was founded as an accrediting body to develop common standards for the exchange of data between European member states. The European DNA Profiling Group (EDNAP) has conducted a quality assurance which in Germany is carried out annually by intermultiple laboratories testing from the German DNA Profiling Group (GEDNAP). These interlaboratory comparisons present a proper quality control and have demonstrated that excellent reproducibility can be seen between forensic

[1] Exceptions are special cases such as identical twins, transplant recipients or chimeric individuals, which are not discussed here.

laboratories [50]. Furthermore, the collaborative work between different laboratories as well as police institutions on a national or international level requires a consistent nomenclature. Therefore, several guidelines for the nomenclature of STR makers were made to avoid redundant designations [58, 59]: STR markers are named after their position in the genome, and markers which are located inside a gene carry the name of the respective gene. For example, TH01 is in the first intron of the gene tyrosine hydroxylase. Sequential numbers are used to give uniqueness to other markers located outside of genes. For example D16S539 is named after: **D** = DNA, **16** = chromosome 16, **S** = single copy sequence, **539** = 539th locus described on chromosome 16 [50].

Although the data analysis of STR alleles from electropherograms is relatively straightforward, it does require considerable expertise by the scientist. During CE several biological and technological related artifacts can create extra peaks and therefore significantly complicate data interpretation, especially for samples with poor quality [60] (Figure 1-3). Technology-related artifacts can for instance be dye blobs and spikes. Dye blobs typically are created by free fluorescent dyes which have failed to be connected to a primer [61]. Spikes are in general sharp peaks that can arise through the presence of air bubbles or urea crystals during electrophoresis [62]. Like any analytical instrument also CE produces a baseline "noise" which can contain additional small peaks. Therefore, peak detection thresholds are set on the instrument and signals below this detection or analytical threshold are considered as unreliable. A common peak detection threshold is 50 relative fluorescence units (RFUs).

On the other hand, biology-related artifact peaks occur for instance through incomplete 3'(+A) adenylation [63]. By its nature the *Taq* polymerase normally catalyzes the addition of an extra adenine on the 3'-end of the double-stranded PCR products. An incomplete adenylation however results in an extra peak and typically occurs when PCR was performed with either too much DNA template or under non-optimal thermal cycling conditions. Extra peaks caused by incomplete adenylation can be especial then critical for data interpretation when they occur in true allele positions as 11.3 or 19.1. Furthermore, point mutations or a duplication can result in additional primer binding sites and cause triallelic patterns [64].

The most common biological artifacts are stutter products which are formed during PCR by slippage of the polymerase [65-68]. Stutter artifacts therefore most often exhibit one repeat less or more than the corresponding main allele (e.g. n-4 or n+4 for tetranucleotide repeats). Because stutter artifacts are difficult to avoid, each laboratory has to develop stutter thresholds to assign an extra peak in a stutter position either as being due to stutter or as a real allele. Usually stutter thresholds are selected to be < 15% of the main peak but thresholds can vary between different loci. In LCN

samples it often occurs that true alleles are masked by stutter. Finally, extra peaks can also result from a mixed DNA sample containing more than one contributor.

In addition to extra or artificial peaks, allelic dropout can also complicate data interpretation. In the majority of cases, dropouts are caused by unsymmetrical PCR amplification of the tow alleles, which in turn is most likely caused by poor DNA quality or low concentration. However, allelic dropout also can arise through an overload of the PCR with DNA or through unexpected sequence variation at the primer binding sites (null alleles). Forensic laboratories are therefore expected to use an experimentally determined dropout threshold for the identification of homozygous alleles [69]. The proper use of dropout and stutter thresholds becomes especially important for the analysis of mixtures, which is probably the most difficult case of data interpretation.

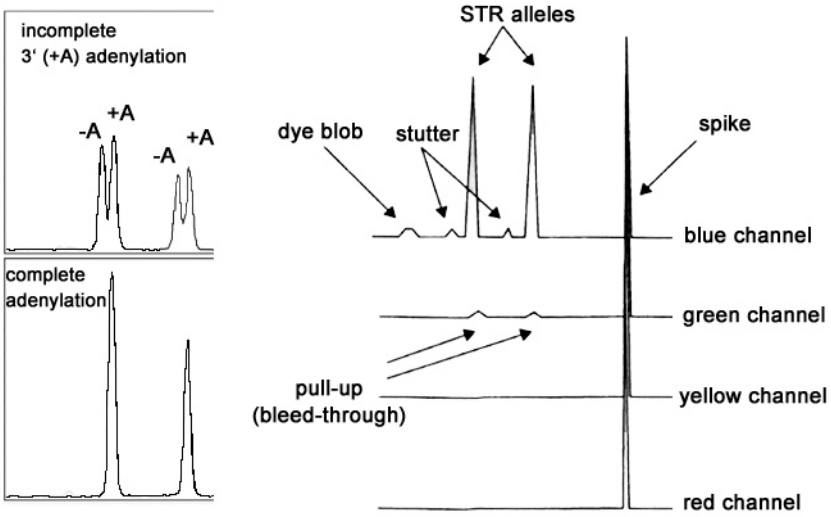


Figure 1-3: Common types of artifacts during STR analysis.
The figure is redrawn from Butler 2009 [50] and the AmpFlSTR® Yfiler™ manual [70].

1.3 Resolution of mixed STR profiles

In forensic work, DNA mixtures arise when two or more individuals contribute to the same DNA sample which can occur in a number of evidentiary situations, such as fingernail clippings or swabs taken from the skin or body orifices, from which it is usually difficult to separate the different cell types. The presence of a DNA mixture can be identified by the occurrence of additional peaks. The detection of three or four peaks within one STR marker for example indicates the presence of at least a second contributor. Peak height imbalances between heterozygous alleles and abnormally high

stutter products are also indications for a DNA mixture. Today's ability to detect a minor component in the DNA profile of mixed samples has improved dramatically due to very sensitive detection technologies (PCR coupled with fluorescent measurements) and an increasing number of markers. The advantage of PCR coupled with fluorescence-based detection is besides its high sensitivity the ability to track relative DNA quantities from multiple contributors in the form of consistently different peak heights and areas that correspond to each respective sample. Therefore, peak information, in particular the correlation between multiple peaks occurring at a similar level versus a second level of others, can reliably be used to determine minor and major components of the mixture and to decipher the possible genotypes of the contributors.

However, extra peaks first have to be convincingly identified as true alleles, i.e. arisen by an additional DNA component and not due to biological or technical artifacts as discussed above in chapter 1.2 [71]. The minor component of a mixture is usually detectable down to mixture ratios of 1:10 or 1:20, depending of the sensitivity of the used PCR [54, 70, 72, 73]. The interpretation of a mixed profile can be very challenging, depending of the type of mixture. The German Stain Commission *(Spurenkommission)* categorizes mixed profiles into several types. The ones best to interpret are mixtures that have clearly distinguishable major and minor DNA component ratios of 1:3, 1:4 or less [74, 75]. However, most of the mixtures obtained in practice display mixed ratios less than 1:3, with a major and minor component that is not clearly distinguishable.

Mixture ratios often cannot be accurately calculated at every locus with complete confidence, especially if two or three peaks have shared alleles. Gill *et al* 2008 [76] illustrate the challenges of mixture interpretation for several examples: The first example (Figure 1-4 A) assumes a two-person mixture with the peaks A and E as being minor contributors, and not located in stutter positions. Since peak B is below the stutter threshold it can therefore be unambiguously designated as a stutter and discounted from interpretation. In the second example (Figure 1-4 B), allele A is again above peak B, which is below the stutter threshold, but here no additional peak E occurs. If the peaks C and D show sufficient heterozygous balance (meaning both peaks show similar peak heights), then the minor component can be assigned as AA. However if the allelic combination CD is unbalanced, then it may be necessary to include AC or AD as potential minor contributors. In a third example, (Figure 1-4 C) the peaks A and B are below stutter threshold. The low A peak indicate that a second contributor could be present to only a small fraction. In this case the B peak may be the second allele of the minor component (contributor is AB) or it may be a stutter (meaning the minor contributor is AA). Because of the low intensity of the minor component also the composition AF, with F as an allelic drop out and B as a stutter, should be considered.

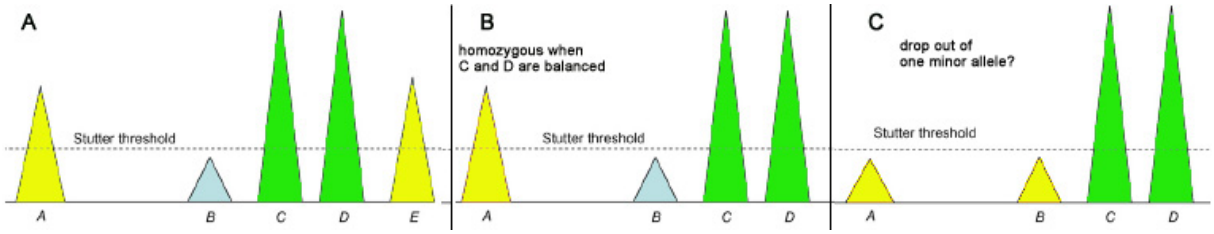


Figure 1-4: Theoretical examples of two person mixtures.
(A) Mixture with major peaks CD (green), minor peaks AE (yellow) and an additional peak B in stutter position (blue). (B) Mixture with major peaks C and D, a minor peak A and an additional peak B in a stutter position. (C) Mixture with the major peaks CD and the minor peak AB, whereas B is in the stutter position. Note that these illustrated theoretical scenarios do not include possible allele designation, which can occur when peaks lie under the drop out threshold. The figure is redrawn from Gill *et al.* 2008 [76].

Accurate mixture analysis is a subject to an ongoing debate. Several guidelines, probabilistic expert systems and biostatistical approaches for resolving complex DNA mixtures have been developed and are discussed among the forensic community [77-81]. The International Society of Forensic Genetics (ISFG) for example has released recommendations on the interpretation of mixtures that discuss two general approaches: the probability of exclusion (PE) and the likelihood ratio (LR) approach [82].

The PE approach provides an estimate of the portion of the population which has a genotype composed of at least one allele not observed in the mixed profile [50]. The advantages of the PE approach are that it requires no prior knowledge about the number of possible contributors to a mixture, and it is easier to explicate in court. However, a PE calculation does not make use of all available information, such as peak heights and positions. The PE approach is therefore restricted to unambiguous DNA samples. If for example the DNA profile shows some minor alleles that happen to have the same size as stutters, and/or if allelic drop out is possible, then PE does not make full use of the evidence and is problematic to apply. In this case an LR approach is preferable.

The likelihood ratio (LR) approach is the ratio of possibilities under alternative propositions and provides a reliable method that is able to make full use of available genetic data [75, 83]. The advantage of the LR calculation is that stutter, heterozygous balance and drop out can be assessed probabilistically. However, LR calculations can be very complex and conclusions are not necessarily straightforward – which can present a considerable problem in court. Moreover, depending on the circumstances of the case, LR calculation can be hampered, because it involves assumptions about the number of contributors to the mixture [84].

The main critical issue for biostatistical calculations of mixed profiles however is the fact that they may result in highly varied appearances depending on the quality and quantity of sample, the number of contributors and the circumstances of the case. For example, profiles can be inconsistent and show some loci where potential multiple contributors can be statistically treated as originating from a single source, while other loci may be too complex to confidently attribute alleles. Here

statistical approaches can help to determine a partial profile (= only a fraction of markers from one multiplex PCR is given). The accuracy of these partial profiles might still be questionable, because of stochastic effects that may occur during PCR. The use of replicates and different PCR kits are therefore highly recommended to overcome possible bias and statistical effects but in practice they work only in part [85, 86]. The most frustrating case however is when the samples show high quality and complete profiles, but are composed of a 1:1 mixture ratio and therefore make it impossible to generate a profile for a search in the forensic DNA database. In such cases today's methods unfortunately do not allow the resolution of mixed DNA samples of the contributors.

One alternative method described to solve complex DNA mixtures is the use of a vast set of single nucleotide polymorphisms (SNPs) as for example described by Voskoboinik and Darvasi, who used a panel of 1000 to 3000 SNPs [87]. However, despite it could be shown that SNP analysis present an possible alternative, the high cost and labor involved in genotyping thousands of SNPs makes it currently not applicable for routine forensic practice. Moreover, only little allele frequency data exists for an appropriate reference population with almost no experience on how to deal with genotyping errors.

Hall and Castella developed another interesting assay to resolve mixed samples which they call STR-DIP assay [88]. In this approach, they use PCR primers that are specific to insertion-deletion polymorphisms (DIPs) located in close proximity to forensic STR markers. In mixed samples, where both contributors differ in one DIP near a forensic STR marker, these specific primers can be used to amplify only one of the contributor allele. In cases like sexual assaults the minor DNA component, often belonging to the perpetrator, frequently is under the 10% threshold and therefore masked by the major component. Here the specific PCR of the STR-DIP assay allows the rescue of the minor component alleles and also might help to solve complicated allele overlaps. However, in the case of balanced mixtures this assay fails because it is unable to connect different markers to one individual profile.

Further approaches, such as differential lysis or microdissection of single cells aim the physical separation of cells and therefore to avoid the formation of a DNA mixture. They have been mostly applied in rape case, where sperm cells of the offender could be isolated but the efficiency of separation and DNA yield of the specific cells is limited [57].

At present, the most reliable and common tool for the recovery of a minor male DNA component from female/male mixtures is the use of Y-chromosome specific marker and will be described in more detail in the following chapter.

1.4 The human Y-chromosome

Among all chromosomes of the genome, the Y-chromosome plays a separate role. M. Jobling and C. Tayler Smith describe very colourful its properties as a: " list of violations of the ruled book of human genetics": it does not recombine, it is a genetic wasteland with only few genes, but massive repeated satellite DNA and it is not required for life (females do well without it) [89]. This description sounds almost offending, but just this disregard of the rules makes the Y-chromosome to an excellent tool for the study of genomic mechanisms, human evolution, and because of its male specificity it is also attractive for forensic genetics. The highly repetitive landscape of the Y-chromosome makes it markedly hard to resolve, and the Y chromosomes of only three species have been completely sequenced – those of human, chimpanzee, and rhesus macaque [90]. The most prominent role of the Y-chromosome is the determination of the sex and the production of sperms. Hereby the region described first, is the SRY (sex-determining region Y) which is expressed early during embryonic stage and initiates testis differentiation. In total the Y-chromosome holds 78 transcribed genes which belong to 27 protein types. Many of them show several duplications for example of genes that lie inside the AZF (azoospermia factor) regions. For example the AZFc region comprises a 4.4 Mb sequence and 3 massive palindromes. De novo deletions of the AZFc arise in roughly 1 of 4000 males and causes nonobstructive azoospermia (no sperm detected in semen) or oligozoospermia (sperm count less than 5 million/ml) [91]. The organisation of the Y-chromosome can be categorized based on two different perspectives: an evolutionary perspective, which compares the Y-chromosomal to the X-chromosomal sequence, or a structure-centered perspective, which looks for blocks of sequences with different composition. The evolutionary perspective tries to explain the origin of the Y-chromosome from an identical pair of ancient autosomes. Hereby, it is proposed that the proto-Y-chromosome developed approximately 300 million years ago (Mya) through an inversion of a huge block of sequence [92]. This inverted region of the Y-chromosome was then not anymore able to undergo meiotic recombination with the X-chromosome and started to differentiate on its own [93]. Without recombination however, ancient genes cannot be maintained and consequently suffer from degeneration and gene loss.

The comparison of several Y-X homologues showed, that - over the course of hundreds of millions of years - at least five huge inversion events occurred on the Y-chromosome, with each inversion further suppressing the likelihood of successful recombination with the X chromosome. Y-chromosomal evolution therefore went on in stages, which can be seen as evolutionary strata on the Y-chromosome sequence. Every stratum shows an increasing similarity between X-Y homologues, with more recent ones becoming inversion events. The youngest inversion event is supposed to have

been occurred approximately 25 Mya, forming a stratum with inversion breakpoints that are still recognizable [94].

Next to Y-chromosomal inversions the modern Y-chromosome was shaped also by large X to Y transpositions. The youngest proof for an X to Y transposition indicates a Y-chromosomal sequence on the short arm which is 99% identical to DNA sequences in Xq21. This Y-chromosomal sequence has a combined length of 3.4 Mb and was transferred about 3-4 million years ago, after the divergence of the human and chimpanzee lineages [95-97]. Because of its origin, this sequence is called X-transposed and represents one of several defined structural (or functional) classes into which the Y-chromosome can be classified (Figure 1-5): First of all, the Y-chromosome can be distinguished based on its sequence-dependent ability to still recombine with the X-chromosome. Hereby, the pseudoautosomal regions (PARs), which comprise only 3 Mb of the entire 60 Mb of the Y-chromosome, are still able to recombine with the X-chromosome and are located at both ends of the Y-chromosome. The major part of the Y-chromosome belongs to the non-recombining region (NRY) or also called the male-specific region (MSY), and is isolated from recombination. The MSY can be divided in heterochromatic and euchromatic sequences. The latter contains three classes which are called X-transposed (described above), X-degenerated, and ampliconic (Figure 1-5).

In contrast to the X-transposed sequence blocks, the X-degenerated segments have less sequence similarity to the X-chromosome (between 60 and 96%) and are thought to be survival relics of ancient autosomes from which the X- and Y-chromosomes co-evolved [92]. The amplicon sequences are highly repetitive with long stretches of duplications and are most often arranged in palindrome structures that house multicopy gene families with testis-specific expression, like the AZF region. Although the structure of the Y-chromosome itself only plays a subordinate role in forensic genetics, scientists have got to be aware of it in order to successfully design experiments for Y-chromosomal markers. The high similarity between large sections of the Y- and the X-chromosome brings with it the inherent danger of unspecific primer design. Furthermore, the highly repetitive structure as well as the duplications inside the Y-chromosome can lead to multicopy markers which can complicate the analysis. Additionally, due to ethical considerations markers, located inside known genes have to be avoided.

Only one Y-chromosomal gene, the amelogenin Y (AMELY), is used for forensic analysis as a sex determining locus. The AMELY is 3272 bp long and located on the 11q12.2 region which belongs to the X-degenerated sequence. A gene homologous to AMELY, the AMELX, is located on the p22 region of the X-chromosome and has a size of 2872pb [98]. Both genes are expressed in teeth, whereas mutations of the AMELX cause so called X-linked amelogenesis imperfecta, a disorder of the formation of the teeth's enamel [92, 99]. The most commonly used amelogenin PCR-based sex test

is the one described by Sullivan *et al.* 1993 [100], in which primers flank a 6pb deletion within the first intron of the homologues genes [101]. Also modern profiling systems, such as the AmpFSTR typing kits from Life Technologies / Applied Biosystems or the PowerPlex kits from Promega, contain those AMEL primers [102].

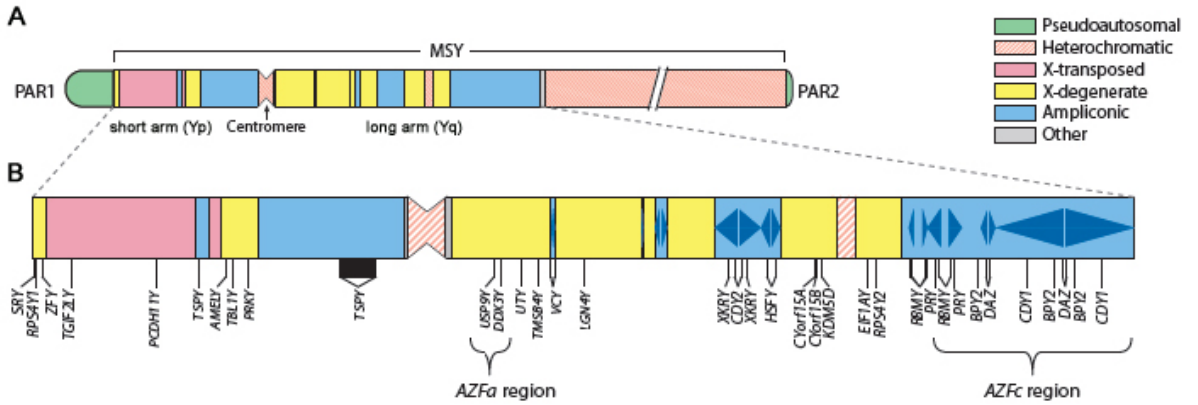


Figure 1-5: Structure of the human Y chromosome.
(A) The Y-chromosome consists of two pseudoautosomal sequences (PAR1 - 2.6 Mb and PAR2 – 0.5 Mb [51]) at each end of the Y chromosome and a male specific region (MSY). (B) Detailed presentation of the three different classes of the euchromatin of the MSY, without the 30 Mb heterochromatin. The MSY comprises 8 Mb on the short arm and 14,5 Mb on the long arm [92]. Positions of the protein coding genes are indicated by vertical lines. The scheme is redrawn from Hughes *et al.* 2012[90].

The Y-chromosome offers for forensic investigations much more than only sex determination. In the last two decades, the Y-chromosome became important for a number of different forensic applications such as paternity testing, solving of rape cases, sexual assaults, missing persons, unknown military grave and mass disaster investigations, human migration and genealogical research [50]. All of these applications benefit from the fact that the biggest part of the Y chromosome, the MSY (approx. 95%), is transferred as an identical copy from the father to the son. In other words - and in contrast to autosomal STR markers - none of the Y-chromosomal markers are shuffled during meiosis but instead are inherited as a single haplotype block. In 1997, the European forensic community settled on a core set of Y-STR markers, or a “minimal haplotype”, that includes DYS19, DYS389I/II, DYS390, DYS391, DYS392, DYS393 and DYS38a/b. In the following years the minimal haplotype was extended several times, whereas currently available typing kits have increased the number of markers up to 23 (Table 1-2). These Y-chromosomal STR markers also display variable number of repeats and are analysed in the same way as the autosomal STR systems described above (chapter 1.2).

Table 1-2: Set of loci included in commercially available kits.

Haplotype name	n	Typed loci	PowerPlex Y® (2003 Promega)	Mentype® Argus Y-MHQS (2008 Biotype)	AmpFLSTR® Yfiler® (2006 Applied Biosystems)	PowerPlex® Y23 (2012 Promega)
Minimal[a]	9	DYS19,DYS389I+II, DYS390, DYS391, DYS392, DYS393, DYS385a/b	x	x	x	x
SWGDAM[b]	11	Minimal + DYS438, DYS439	x		x	x
Powerplex[c]	12	SWGDAM + DYS437			x	x
Yfiler[d]	17	Powerplex + DYS448, DYS456, DYS458, DYS635, YGATAH4			x	x
Powerplex Y23[e]	23	Yfiler + DYS481,DYS533, DYS549, DYS570, DYS576, DYS643				x

(a) First evaluation of Y-chromosomal STR markers [103]. (b) In 2003 the U.S. Scientific Working Group on DNA Analysis Methods (SWGDAM) recommended a core set of Y-STRs which include the minimal haplotype and 2 additional markers. (c) The PowerPlex incorporates the European minimal haplotype, the recommendation of the SWGDAM and one additional marker [104, 105]. (d) Extension of the Y-STR haplotype [70, 106]. (e) Incorporation of rapidly mutating Y-STRs [107, 108]. The dates (years) provide the time of launching. n = number of loci. Table is redrawn from Willuweit 2010 [109].

The determination of Y-chromosome markers and haplotypes is also broadly applied for the resolution of mixed profiles. In crime cases, such as sexual assaults, vaginal swabs from raped or murdered female victims often show a mixture with a major female and a minor male component. Autosomal analysis of these mixed stains often fail in amplifying the minor male component because of the domination of female DNA [110]. However, the occurrence of a Y-chromosome only in males allows for a very sensitive detection of the Y-STR markers, even in the presence of an extremely high background of female DNA: Tests performed on samples artificially set up with male to female DNA ratios of 1:2000 still gave optimal typing results of the Y-STR markers [70, 111].

The group of Prinz *et al*. 1997 [111] showed between the years 1998 to 2001, that 13% of their sexual assault cases contained DNA of multiple semen donors. Considering this context, it has been demonstrated that Y-chromosomal analyses of male/male mixtures provide the advantage of being less complex due to the nature of their haplotype. The typing of only one allele per STR locus greatly facilitates data interpretation and also gives more confidence in the number of contributors. On the other hand, although Y-chromosomal haplotypes are an excellent tool for the exclusion of suspects, they do not display the same power of discrimination as autosomal STR markers. Because of the missing recombination of the MSY, the Y-loci are not independently inherited and therefore the probability for a random match cannot be estimated based on the product rule (chapter 1.2). Therefore the probability of inclusion has to be calculated on the basis of the haplotype frequency alone instead of through the genotype frequencies.

At the present time, there are two approaches to calculate match probabilities for Y-STR haplotypes [110, 112]: The counting method and the haplotype surveying method [113, 114]. For both methods, large, representative and quality-assessed databases of haplotype samples from appropriate reference populations are required [115, 116]. Hereby the counting method simply counts the number of matches of one Y-STR haplotype in the databases. The obtained frequencies are then corrected for database size and sampling errors by adding a 95% confidence interval. In cases where the requested haplotype cannot be observed in the database, its frequency can be reported as the upper confidence interval limit or by adding the unobserved haplotype to the database (1/database +1) [116, 117]. However, in cases of non-observation the counting method may reduce the evidentiary power. Here, Roewer and colleagues recommend the use of an extrapolation method which they call “frequency of surveying”. The surveying method is based on similarity and frequency relationships and has the advantage that it does not require the presence of the searched haplotype in the reference database. As a consequence, also rare haplotypes retain their high evidentiary power, even when the database used for the estimation is of moderated size [114]. It is therefore critically important that reference databases use appropriate quality assurance procedures to ensure that only representative and informative data is being used so as to display a most realistic distribution of haplotypes [115]. Currently, a number of online databases exist, most of which associate haplotype information with family surnames rather than with haplotype frequencies, e.g. the Sorenson Molecular Genealogy Foundation or Ysearch from the FamilyTree DNA service. Database frequency estimations are available from the US Y-STR database (US Y-STR) and the Y-STR Haplotype Reference Database (YHRD). The latter comprise at present 112005 samples from 834 populations (Release 43 build, 2013). The YHRD also offers a “mixture tool”, that calculates a likelihood ratio for donorship of a know profile versus non-donorship – but it is limited to haplotypes which do occur in the database [113, 118]. Next to forensic applications, Y-chromosomal STR databases are important tools for the study of human migration and evolution from a male perspective. The inheritance of unchanged Y-chromosome haplotypes over generations enables scientist to follow back male lineages, just as mitochondrial DNA is being used to trace female lineages. Genetic drift and isolation, for example caused by migration, territorial separation or social barriers, lead to partial genetic differentiation and the formation of subpopulations with different haplotype frequencies. Hereby the relatively high mutation rates of STRs also allow for the examination of quite recent demographic events, and even give insight into population structures and histories in areas with high gene flows, like for example Europe and the Fertile Crescent, which have been ground stages for migration and invasions.

In a study with 12700 European haplotypes of the YHRD database, Roewer and colleagues showed a major genetic division of European males into Slavic-speaking eastern and Romanic language-speaking western populations [119]. The group of Kayser could distinguish between Polish and German populations by the comparison of 913 Polish and 1212 German haplotypes [120]. At present, there exist innumerable population studies, all contributing to shape our understanding of human evolution and prehistoric and non-recorded history. Many of these studies also include the examination of Y-chromosomal single nucleotide polymorphisms (SNPs), which display much lower mutation rates of about 1×10^{-9} per locus per generation. The relatively high mutation rate of the STR systems makes it difficult to identify ascertain observed haplotype as being either modern or ancient, or, in other words, to reliably differentiate between "identity-by-state" and "identity-by-descent" [121]. In comparison, the identification of the ancestral versus the derived state of a certain SNP is remarkably much easier, since SNPs are generally binucleotide polymorphisms and very rarely undergo back-mutations that would confound the analysis. Furthermore, Y-chromosomal SNPs represent a valuable source of phylogenetic markers which very well reflect pioneering migration and ancient colonization (see more details in chapter 1.51.5).

However, the advantage of the Y-chromosome as a male-specific lineage marker is at the same time also its greatest limitation. The inheritance of identical haplotypes from father to son has two unfavorable consequences: first, as described above, match probabilities depend on database size and are in general reduced. Second, any matches between a suspect and the available DNA evidence also include all male relatives from the same paternal lineage, such as brothers, uncles and even distant cousins [122]. Y-chromosomal markers therefore provide less evidentiary power than autosomal STR profiles and on court. To improve the differentiation between male relatives, Manfred Kayser and colleagues have introduced several rapidly mutating Y-STRs with mutation rates above 1×10^{-2} [123]. On average the mutation rates for these markers are one order of magnitude greater than those of ordinary STR markers (1×10^{-3}) [124, 125]. The PowerPlex® Y23, which has been launched in July 2012 from Promega, includes already two of the rapidly mutating Y-STR markers and extends the AmpFLSTR® Yfiler by six further STR markers in total (Table 1-2) [107].

The extension of commonly used Y-STR sets would also improve the resolution of male lineages in defined populations [124, 126, 127]. The use of only a small set of Y-STR markers holds the risk of finding the same haplotypes for even non-related males, as it was shown from Vermeulen *et al.* 2009 [128]. His group showed that at least 25 highly informative Y-STRs were necessary to reach the highest possible male lineage differentiation in a global sample set of 590 males from 51 populations. However, further additional markers might be necessary for the complete resolution of a population

that is characterized by low internal genetic structure and a demographic history that may include long periods of isolation. The extension of further haplotype markers therefore is in general a desirable development and triggered by constantly increasing technology and company interests.

1.5 SNP classes and databases

Single nucleotide polymorphisms (SNPs) represent the simplest and most abundant form of genetic variation in the genome. However, in the field of forensic genetics they never – or at least not yet - reached the same importance as their counter-parts, the STR markers. Reasons for their reduced significance are that SNPs are typically more complicated to analyze and less informative, since most of them are only biallelic. The advantage of SNPs is that their analysis requires only small amplicon sizes, typical less than 100 bp. Therefore SNP-based assays show higher success rates for degraded DNA samples. There have been only a few attempts to develop identity testing SNP panels [129]. Major work was done by the group of Kidd and colleagues which screened hundreds of candidate SNPs in 44 populations. They ended up with a set of 86 SNPs without significant linkage disequilibrium and suitable heterozygosities. Finally they presented a minimal set of 45 SNPs with which they were able to obtain probabilities for individual genotypes of less than 10^{-15} [130, 131].

In addition to SNPs used for identity-testing, there are three SNP classed related to forensics: Lineage-informative SNPs, ancestry-informative SNPs and phenotype-informative SNPs. The latter gained much attention over the last years by associating certain SNPs to specific phenotypes like the eye color of a person. In 2011 Manfred Kayser and colleagues presented a multiplex kit predicting eye color, the IrisPlex, which is based on the detection of 6 eye color-informative SNPs. The data from the genotyped SNPs allows, through the use of a complex regression model, the prediction of blue or brown eye color with 90% accuracy [132-134]. Nowadays there are several projects aiming to improve the prediction success of also intermediate eye colors, as well as for hair and skin color and even for the estimation of size and age of a person at the time when its DNA sample was left at the crime scene [135-137].

Other than phenotype-related SNPs, ancestry-associated SNPs have their main importance in the research of human migration and evolution. Hereby the major sources of ancestral-informative SNPs are uniparentally inherited marker systems from the mitochondrial or Y-chromosomal DNA, which allows the determination of a maternal or paternal hierarchic descent order, respectively. In the past 25 years, hundreds of SNPs have been investigated and categorized into defined haplogroups and built up mitochondrial and Y-chromosomal phylogenetic trees with detailed genealogical branching

order [138]. Pioneering work in establishing the Y-chromosomal tree was done by the Y-chromosome consortium (YCC), who published the first tree in 2002 and developed a hierarchical nomenclature system (Figure 1-6) [139]. Hereby all branches or haplogroups define patrilineal lineages, which can be followed back to a most recent ancestor, also called the Y-chromosomal Adam, a theoretical male who lived about 140kya in Africa [140, 141]. Worldwide studies of phylogenetic SNPs resulted in the setup of 20 major clades and gave detailed information about the origin of haplogroups and the Y-chromosome diversity [142]. For example the oldest clades (haplogroups A and B) of the Y-chromosome tree have been found exclusively in Africans and therefore strengthened the "Out of Africa" hypothesis of human evolution. Younger branches like R1 and its sub-branches describe the settlement of Europe. The distribution of haplotype J fits archaeological data for the introduction of farming from the southeast [89].

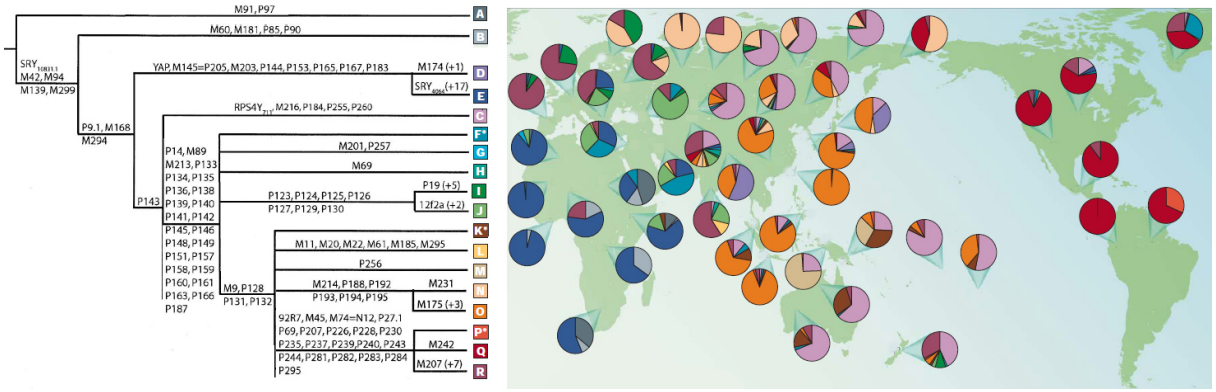


Figure 1-6: The Y-chromosomal parsimony tree.
The figure shows a simplified form of the Y-chromosomal phylogenetic tree with its major clades A-R (left side) and their global distribution (right side). Names of the phylogenetic SNPs are given along the branches, which they define. For example the derived state of the SNPs M91 and P97 define haplogroup A, which occurs mainly in central and South Africa. Figures are adapted from Karafet *et al.* 2008 [142] and Jobling and Tyler Smith 2003 [89].

Although the phylogenetic SNPs are primarily of interest for studies of human history and mutation process, they can also be beneficial in forensic investigations. Y-chromosomal haplogroups as well as Y-chromosomal STR haplotypes show an extreme geographically-correlated structure and therefore correspond to orthographic information, which can be helpful in tracing the paternal population of origin [110]. In cases where nothing is known about the perpetrator, the biogeography ancestry of an individual may indirectly provide some limited information [129]. Y-chromosomal SNPs can also be helpful to analyses of contact traces during sexual assaults which result only from mere touching during the sexual assault or the attempt of rape. DNA extraction of these contact traces result often in only degraded DNA samples, in which the minor male components of the contact traces are often undetectable by Y-chromosomal STR typing (see chapter 1.3).

The last category, the linage markers, describe, in contrast to the ancestry-informative markers, uniparentally inherited SNPs which are not included in any phylogenetic tree. Therefore these markers can show very low frequencies or be even family specific and might also be defined as private mutations. In addition to phylogenetic SNPs, lineage-informative SNPs can also be used for kinship testing and family tree studies [129].

All the SNP classes discussed here are available in several databases. Sets of the most well-established and validated SNPs are based on the SNP collections of the phylogenetic trees. Other databases, like the HapMap database, are specialized in collecting SNPs for medical usage, where SNPs are used to map genes which are involved in monogenetic or complex disorders. One of the biggest collections of SNPs is the short genetic variation database (dbSNP) of the National Center for Biotechnology Information (NCBI) which, for example, includes also lineage or identity testing SNPs described above.

The availability of high-throughput, cost effective next-generation sequencing and other methods for detecting SNPs has caused rapidly growing numbers of new public database entries [60, 143, 144]. This is particularly true for the NCBI database, where the number of SNP entries increased rapidly after the release of the human genome in April 2003; at present, there are 53.567.890 SNP entries and over 100 million submissions (NCBI database, Build 137). Although these databases represent a powerful scientific tool, they have not overcome the problem of substantial error rates. Once a false SNP is placed in the database, the error becomes invisible and remains. Several attempts have been made to determine actual error rates and verify SNP databases. For example, Reich *et al.* 2003 [145] sequenced 173 kb across 17 loci in 150 chromosomes from individuals of European and West African ancestry and estimated an error rate of 12%. In the same year, Carlson *et al.* 2003 [146] sequenced 50 genomic regions that spanned 564 kb in 47 European and African American individuals and calculated an error rate of 35-36%. These and other works have uncovered a vast number of false entries and provided the first assessments of the quality of SNP databases [147].

Furthermore, it has been shown that a proportion of so-called SNPs actually were computer-generated nucleotide mismatches from paralogous copies of duplicated sequences and should be more appropriately labeled as paralogous sequence variants (PSVs) [148-150]. In 2006, Musumeci and colleagues estimated the rate of PSVs in the dbSNP database to be 8.32% [151]. In addition, most databases contain a substantial fraction of nonpolymorphic SNPs with allele frequencies below 0.02 [152] which in this work will be described as private SNPs.

1.6 Aim of this study

The aim of this study was to develop a Y-chromosome specific extraction method which physically separates individual Y-chromosomal STR haplotypes from male DNA mixtures and therefore recovers individual marker information from otherwise non-interpretable mixed profiles. Therefore, the technology of the haplotype-specific extraction method was adapted for the use of Y-chromosomal extraction probes and the final downstream processing of STR genotyping by multiplex PCR and capillar electrophoresis. The goal was to establish allele-specific extraction probes and to optimize the Y-chromosomal HSE approach to such an extent that the enrichment of the Y-STR markers of one contributor is sufficient for a complete separation effect in subsequent data-analysis.

2 MATERIAL AND METHODS

2.1 Instruments and software

- **Centrifuges:**
  - Multifuge (Heraeus, Berlin, Germany)
  - Centrifuge 5415 (Eppendorf, Hamburg, Germany)
  - Speedvac SC100 (Savant Instruments Inc., Farmiongdale NY, USA)
  - Micro-centrifuge for strips and single tubes (Kisker Biotech GmbH, Steinfurt, Germany)

- **DNA extraction robots and equipment (Qiagen, Hilden, Germany):**
  - BioRobot® EZ1 Workstation
  - BioRobot® M48 Workstation with Filter Tips 1000 µl, reagent containers small and large, sample Prep plates and software GenoM48 v.2.0

- **DNA quantification:**
  - Quarz-cuvette SUPRASIL®, type 105.210-QS, light path 5 mm, centre 15mm (Hellma GmbH & Co. KG, Müllheim, Germany)
  - Rotor-Gene Q and software 2.0.2.4 (Qiagen, Hilden, Germany)
  - Spectrophotometer GeneQuant (Amersham Bioscience, Freiburg, Germany Biochrom Ltd., Cambridge, England)

- Electrophoresis – chamber (MWG-Biotech, Ebersberg, Germany)
- Magnetic particle concentrator Dynal MPC-S, tube rack for 1,5 and 2 ml tubes (Life Technologies, Darmstadt, Germany)
- Magnetic stirrer MR 2002 and magnetic stir bars (Heidolph, Schwabach, Germany)
- Pipettes 0.5-10 µl, 10-100 µl and 100-1000 µl (Eppendorf, Hamburg, Germany)
- Power supply (Renner GmbH, Dannstadt, Germany)
- Primer design software Oligo Explorer 1.2 (http://www.genelink.com)
- Primer design software OligoAnalyzer 1.1.2 (http://oligo-analyzer.software.informer.com)
- QIAxel (Qiagen, Hilden, Germany)

- **Pyrosequencing (Biotage, Uppsala, Sweden):**
  - PSQ 96MA robot and the software PSQ 96MA 2.1.1
  - Pyrosequencing Sample Prep Tool and magnetic sticks

- Pyro-Assay Design software 1.0

- Scale (Sartorius, Göttingen, Germany)

- **Sequencer and equipment** (**Applied Biosystems, Foster city, USA):**
  - Sequencer ABI Prism® 3130-Avant Genetic Analyzer
  - Capillary Array, 16x36 cm capillaries,
  - 3130+3100 series Plate Retainer 96 well,
  - 3130+3100 series Plate Base 96 well,
  - Plate septa 96 well,
  - MicroAmp™ Optical 96-Well Reaction Plate
  - Software Sequencing Analysis 5.2
  - Assign ™ DBT 3.5.1 for HLA Sequencing Based Typing (Conexio Genomics, Fremantle, Western Australia)
  - STR-analysis program GeneMapper ID-x1.1.1 (Applied Biosystems, Forster City, Germany)

- Statistic program PASW (IBM, Ehningen, Germany)

- **Thermoblocks and -mixer:**
  - Thermomixer comfort (Eppendorf, Hamburg, Germany)
  - Thermoblock 5436 (Eppendorf, Hamburg, Germany)
  - TruTemp (Robbins Scientific, Sunnyvale, CA, USA)

- **Thermocycler:**
  - DNA Engine Dynal (Biozym Hessisch Oldendorf, Germany)
  - Flexcycler (Analytik Jena, Jena, Germany)

- **Thermometer** (precision ± 0.5°C):
  - calibrated thermometer (LET 0/100 DIN Precision, 5135)
  - electronic thermometer (Testo 1100, 50 - 150°C)

- UV-Illuminator (Hoefer, Heidelberg, Germany)
- Vortex (Jahnke & Kunkel, Staufen, Germany)

2.2 Chemicals and consumables

- DNA Fast Analysis Kit (Qiagen, Hilden, Germany)
- Ethanol 96% (J.T Baker, Deventer, Netherlands)
- EDTA (Pharmacia biotech, Uppsala, Sweden)
- ExoSAP-IT (USB, Cleveland, USA)
- EZ1 DNA Blood 200 µl Kit (Qiagen, Hilden, Germany)

- **Gel electrophorese chemistry:**
  - Agarose A-9539 (Sigma-Aldrich, Steinheim, Germany)
  - Bromphenol blue Na-salt (Serva, Heidelberg, Germany)
  - Ethidium bromide solution 10mg/ml (Sigma-Aldrich, Steinheim, Germany)
  - Ficoll 400 (Serva, Heidelberg, Germany)
  - Flasks 500 ml (Duran Group GmbH, Wertheim/Main, Germany)
  - Glycerol 85% (Herbeta Arzeneimittel, Berlin, Germany)
  - Hyperladder IV, Hyperladder V (Bioline, Luckenwalde, Germany)
  - Nitrile gloves Nitra Tex (Ansell Healthcare Europe, Brussels, Germany)
  - TBE, which contains: 0.445 M Tris-borate pH 8.0 and 10 mM EDTA (Sigma-Aldrich, Steinheim, Germany)
  - Xylene cyanol FF (Serva, Heidelberg, Germany)

- **HSE chemistry:**
  - AmpliTaq® Polymerase (Applied Biosystems, Foster City,USA)
  - dATPs, dTTPs, dGTPs, and dCTPs, 100 mM solutions (USB, Cleveland, Ohio, USA)
  - Biotin-dUTP 1 mM (Roche, Mannheim, Germany).
  - EZ1 HaploPrep Kit (Qiagen, Hilden, Germany)
- illustra MicroSpin™ HR Columns S400 (GE Healthcare, Buckinghamshire, UK)
- Investigator Quantiplex Kit (Qiagen, Hilden, Germany)
- Mag Attract DNA Mini M48 Kit and consumables for extraction with M48 (Qiagen, Hilden, Germany)

- **PCR chemistry:**
  - AmpliTaq Gold® Polymerase (Applied Biosystems, Foster City,USA)
  - GeneAmp® 10x PCR buffer (100 mM Tris-HCl pH 8.3 [at 25°C], 500 mM KCl, 15 mM $MgCl_2$, 0.01% w/v gelatin, Applied Biosystems, Foster City, USA)

- $MgCl_2$ solution (50 mM) (Inviteck, Berlin, Germany)
- dNTPs 10 mM each (Roche, Mannheim, Germany)
- Primer and probes, lyophilized (TIP MOLBIOL, Berlin, Germany and biomers.net GmbH, Ulm, Germany)

- **Pyrosequencing chemistry:**
  - Annealing buffer: 20 mM Tris, pH 7.6, 5 mM magnesium acetate tetrahydrate (Biotage, Uppsala, Sweden)
  - Binding buffer: 10 mM Tris-HCl, pH 7.6, 2 mM NaCl, 1 mM EDTA, 0,1% Tween-20 (Biotage, Uppsala, Sweden)
  - PyroMark Gold Q96 Reagents (5 x 96) (Qiagen, Hilden, Germany)
  - Sodium hydroxide, NaOH (Merck, Darmstadt, Germany)
  - Streptavidin-coated magnetic beads (Biotage, Uppsala, Sweden)

- REPLI-g® DNA Amplification Kit (Qiagen, Hilden, Germany)

- **Sequencer chemistry:**
  - **General Sequencer chemistry (Applied Biosystems, Foster City, USA):**
    - ABI Prism 10x Genetic Analyzer Buffer with EDTA
    - Hi-Di™ Formamide
    - GeneScanTM-600Liz®
    - GeneScanTM-500Liz®
    - Performance Optimized Polymer 4 (POP4)
    - Septa Strips

  - **Chemistry for STR-sequencing (Applied Biosystems, Foster City, USA):**
    - AmpFLSTR® Yfiler, AmpFLSTR® Identifiler plus®, AmpFLSTR® NGM SElect™, and Matrix Dye Set G5

  - **Sanger sequencing chemistry (Applied Biosystems, Foster City, USA):**
    - BigDye™ Terminator v1.1 and v3.1 Matrix Standard Kit
    - BigDye®Terminator (BDT) v1.1 or v1.3 Sequencing Kit (Applied Biosystems, Foster City, USA)
    - BigDye Terminator v1.1 and v1.3 Sequencing Standard Kit
    - Allele SEQR HLA-A, HLA-C, DRB1, and DQB1 (Atria Genetics, San Francisco, USA)

- Sodium acetate - EDTA (Atria Genetics, San Francisco, USA)
- Sodium chloride (NaCl) (Serva, Heidelberg, Germany)

- **Standard consumables:**
  - Aqua ad iniectabilia, sterile water (B.Braun Melsungen AG, Melsungen, Germany)
  - lintless cloth (Drugstore, Berlin, Germany)
  - Parafilm (American National Cam™, Menasha, USA)
  - Tips 10, 100, 1000µl (Eppendorf, Hamburg, Germany)
  - HCl (Carl Roth GmbH, Karlsruhe, Germany)
  - Hygrip gloves latex (Ansell Healthcare Europe, Brussels, Germany)
  - PCR Softstrips, 0.2ml (Biozym, Hessisch Oldendorf, Germany)
  - Sterillium (Bode Chemie, Hamburg, Germany)
  - Tubes 1.5 ml and 2 ml (Eppendorf, Hamburg, Germany)
  - Tubes with screw caps 1.5 ml and 2ml (Sarstedt, Nürnbrecht, Germany)
  - Tubes without lid 2 ml (Sarstedt, Nürnbrecht, Germany)
  - Water (J.T Baker, Deventer, Netherlands)
- Sterile Cotton swabs (Applimed SA, Châtel-Saint-Denis, Switzerland)
- Tris-HCl pH 7.4, 1 M (Sigma-Aldrich, Steinheim, Germany)

2.3 DNA samples

For the determination of Y-chromosomal variation 75 male DNA samples were used from laboratory collections, which covered three ethnic populations and an unknown population group (Table 2-1, Appendix-Table 7-1). Additionally, 20 European female DNA samples have been used for X-chromosomal studies (Appendix-Table 7-2). The DNA for male DNA mixtures was obtained from twelve individual blood samples donated by male laboratory staff (Table 2-8). DNA mixtures for HSE were set up by mixing equal amounts of different genomic male DNAs. The 1:1 ratios of the mixtures were controlled by Y-STR profiling with the use of the AmpFLSTR® Yfiler kit (chapter 2.8.4).

Table 2-1: Laboratory collection of male DNA samples arranged by ethnic group.

	population				
country	**African**	**Eurasian**	**Native American**	**unknown**	
Angola	1				
Cameroon	1				
Egypt	1				
Ghana	3				
Guinea	1				
Ivory Coast	1				
Nigeria	3				
Senegal	1				
Uganda	1				
China		1			
Finland		1			
Germany		20			
Italy		1			
Lebanon		1			
Macedonia		8			
Scotland		1			
Serbia		1			
Turkey		9			
Vietnam		7			
Ecuador			3		
unknown		2		7	
total	13	52	3	7	75

2.4 DNA extraction

Genomic DNA for laboratory collection of male DNA samples was extracted from buccal swabs with the fully automated BioRobot M48 Workstation in combination with the Mag Attract DNA Mini M48 kit. Male DNA samples for HSE testing were extracted with the BioRobot® EZ1 Workstation in combination with the EZ1 DNA Blood 200 µl kit according to the manufacturer's instructions.

These DNA extraction kits include five protocol steps: First the cells are lysed by chaotropic salts. Second, the DNA molecules are bound to the silica surface of magnetic beads. Third, the beads are then washed with ethanol for the removal of the chaotropic salts. Fourth, beads were washed with water for the removal of ethanol. In the last step the DNA is eluted in 50 µl water [153]. For the extraction of DNA from blood samples the extraction robots (M48 or EZ1) were loaded with 200 µl of blood in a 2 ml Sarstedt tube. DNA from buccal swabs were placed in a 2 ml Sarstedt tube and incubated in lysis-buffer (50 mM Tris-HCl pH 7.4, 100 mM EDTA, pH 8.1, 100 mM NaCl, 1% SDS) for at least 30 min at 56°C. After this pre-lyses step 200 µl of the incubated lysis solution is used in the automatic DNA extraction.

Some of the eluted samples still contained traces of magnetic beads remaining from the extraction procedure. To avoid any transfer of silica-coated magnetic beads to the subsequent HSE reaction, the elution samples were placed on a magnetic tube rack to fully separate the DNA sample from any remaining beads. All DNA samples were stored for short term usage at 4°C and for long term usage at -20°C.

2.5 DNA quantification

DNA concentrations of the DNA extracted from 200 µl blood were measured with the Gene Quant spectrophotometer. For the study of HSE based on different DNA input amounts, DNA dilutions of low concentration were prepared and verified by real time PCR.

2.5.1 DNA quantification with the Spectrophotometer

Nucleic acids absorb ultraviolet light at 260 nm with an extinction coefficient of 0.020 $(\mu g/ml)^{-1}$ cm^{-1} for double-stranded DNA and 0.027 $(\mu g/ml)^{-1}$ cm^{-1} for single-stranded DNA. The spectrophotometer measures the absorption at 260 nm and calculates the DNA concentration of the sample according to the Beer–Lambert law:

$$c = \frac{A \, x \, \varepsilon}{b} \qquad (1)$$

Here, A is the absorbance at 260 nm, ε the extinction coefficient and b the path length in cm. The light path of the quartz-cuvette that was used was 5 mm, which allows the instrument to perform exact measurements of DNA concentration between 0.2 - 5 OD_{260} and 10 – 250 ng/µl respectively. As a standard procedure the spectrophotometer also measures the absorption at 280 nm and calculates the A_{260}/A_{280} ratio. A ratio of ~1.8 is generally accepted as indicating “pure” DNA. A ratio appreciably lower than 1.8 indicates the presence of proteins, phenol or insoluble materials, that absorb strongly at or near 280 nm. In this study the A_{260}/A_{280} showed always good ratios because of the automatically DNA extraction with silica beads from material with high DNA content. After a blank measurement DNA samples were quantified by filling 10 µl of sample in a quartz-cuvette, placing it into the instrument and exposing to 260 nm UV light to determine its absorbance.

2.5.2 DNA quantification by Real-time PCR

The quantification of human DNA samples by real-time PCR was performed with the Rotor-Gene Q in combination with the Investigator Quantiplex Kit and the Rotor-Gene software 2.0.2.4 according to manufacturer's instructions. A LED source in the instrument tuned to the absorption maximum of the fluorophore (green channel for absorption at 470 nm) irradiates special sample tubes from the side wall. The chemistry of the Investigator Quantiplex Kit contains human-specific so called scorpion primers for the amplification of 4NS1C®, which is a 146 bp proprietary region that is present on several autosomes of the human genome. Scorpion primers are covalently linked to a sequence-specific probe and contain a fluorophore and a quencher. In an unbound state, the quencher interacts frequently with the fluorophore due to their close proximity thereby suppressing the fluorescence of the probe.

The amplification of the scorpion probe itself during real-time PCR is blocked by a PCR blocking element, like hexethylene glycol, which is attached between the probe and the 5'-end of the primer [154]. After the synthesis of a new strand during PCR, the probe hybridizes with the PCR product and thereby leads to spatial separation of fluorophore and quencher. This greatly increases the ability of the fluorophore to emit light, thereby indicating the successful extension of the scorpion primers and an increase in the concentration of the PCR product. The new fluorescence (emission at 510 nm) is detected by a photomultiplier and the data is sent to a Desktop PC that averages the measured fluorescence of each sample over a number of revolutions.

Additionally, the chemistry of the Quantiplex Kit contains a second primer and a scorpion primer as well as the corresponding template as an internal control for PCR inhibition. The scorpion primer for the internal control uses a different fluorophore ("VIC", yellow channel, excitation at 530 nm and emission at 555 nm). All reaction components for the DNA quantification and the internal control are included in the Reaction Mix FQ and in the Primer Mix IC FQ. Each reaction was set up in 0.1 ml Rotor-Gene Q tubes by mixing 2 µl of sample with 8 µl Reaction Mix FQ and 8 µl Primer Mix IC FQ. Parallel to setting up the samples a serial DNA-dilution (20 ng/µl, 5 ng/µl, 1.25 ng/µl, 0.3125 ng/ul, 0.078 ng/µl, 19.5 pg/µl, and 5pg/µl) was set up as a calibration curve for every new run. All samples and DNA dilutions were run in duplicates.

PCR amplification was carried out with the 72-well rotor with the following parameters: 95°C for 1 min, then 40 cycles of 95°C for 1 s and 60°C for 10 s. For the calculation of the DNA concentration the Rotor-Gene software 2.0.2.4 first sets the "Auto Threshold" to eliminate the background signals. Next, the program calculates the threshold cycle (C_T values) of every sample and the DNA dilution series. The C_T value serves as a tool for calculation of the starting template amount and was defined

as the cycle in which the first detectable significant increase in fluorescence occurs. Final DNA concentration can be than calculated by linear regression of the obtained C_T values of the standard curve. The equation is in the form:

$$y = mx + b \quad (2)$$

where x = log concentration and y = C_T [155].

Each standard curve is evaluated by the calculation of the correlation coefficient (R^2) which has to be between 0.98 and 0.99. Internal control C_T values of about 31 ±1 indicate that no PCR inhibition occurred during the run. In this study the internal control values never indicated PCR inhibition because DNA samples derived from high quality DNA preparations from blood.

2.6 Whole genome amplification

In order to increase DNA material of several samples whole genome amplification (WGA) was performed with the REPLI-g® DNA Amplification Kit. The kit uses Phi29 polymerase and random hexamers for a uniform amplification across the entire genome by multiple displacement amplification [156, 157]. WGA reactions were carried out in 50 µl volume according to the manufacturer's recommendations. First, the template DNA was denatured by adding 5 µl of the alkaline denaturation buffer D1 to 5 µl DNA material and incubated for 3 min at room-temperature. The denaturation step was stopped by adding 10 µl of neutralizing buffer N1. For the amplification, 29 µl REPLI-g Midi Reaction buffer and 1 µl REPLI-g Midi Phi29 polymerase were then added to the denatured DNA. The WGA reactions were incubated at 30°C for 16 hours and afterwards inactivated by heating the sample for 3 min at 65°C. Next, WGA-samples were purified using illustra MicroSpin Columns S400. The columns contain Sephacryl™ resin and allow DNA purification by the process of gel filtration by size exclusion: Molecules larger than the pores of the resin cannot penetrate into the gel and therefore are eluted first. Small molecules however can enter the resin, thus temporarily trapping them in the gel pores and retarding their elution. Columns for each sample were prepared for purification by short vortexing of the resin and centrifugation for 1 min at 2800 rpm. Then WGA-samples were added to the resin of each column and centrifuged again for 2 min at 2800 rpm.

2.7 Database search and selection of SNPs for HSE-probe design

The search of single and multiple- or polynucleotide polymorphisms included the NCBI database, the Mammalian Genotyping Service, the Family Tree DNA and the YHRD database has been used (chapter 7.1). One additional validated polynucleotide polymorphism, MID504, was found in the Mammalian Genotyping Service (Figure 2-1). The following further databases were used for the reconciliation of database entries and for the winning of new information: UCSC Genome Bioinformatics, Human Genome Segmental Duplication Database and Human Organizes Whole genome Database. The selection of SNPs for HSE probe design was based on two criteria: first, on the distance of the SNP to the nearest Y-STR system primary included in the AmpFℓSTR® Yfiler kit and second, on the variability of the SNP within the European population. The existence of the SNPs in the DNA samples was confirmed by Pyro- or Sanger sequencing. In addition to the database entries, two additional SNPs were discovered by Sanger sequencing of a 4 Kb fragment of the DYS437 flanking region (S4:2701 and S6:4204, Appendix-Figure 7-2).

Mammalian Genotyping Service
National Heart, Lung, and Blood Institute

Human Insertion/Deletion Polymorphisms

Marker Name: MID504
Other Marker Name(s):
Type: Diallelic
Chromosome: 25
Other Chromosome: 25
Status: Confirmed
UC-Santa Cruz Sequence Position (05/01/2004)(mb): 3.173533
Other Sequence Position: 2.775294
Allele Length Difference: 3
Gorilla: ND
Bonobo: Same length as short human allele
Forward PCR primer: 5' TGTCACCAGGTACTAGTGCA 3'
Reverse PCR primer: 5' AAAGCAACCACTTTACCTAATG 3'
Expected Human Product Sizes: 132 , 135

Allele Frequencies:

Population	# of Chromosomes Sampled	Long Allele	Short Allele
African	42	0.56	0.43
European	200	0.72	0.27
American_Indian	50	0.73	0.26
Japanese	50	0.70	0.29
Coriell_M44PDR	88	0.74	0.25

Polymorphism Discovery Resource DNA Genotype Data:

PDR Sample #	Allele1	Allele2
1	Long	Long
2	Long	Long
3	Long	Short
4	Long	Long
5	Long	Long
6	Long	Short
7	Long	Short
8	Long	Long

Figure 2-1: Database entry MID504 in Marshfield database of Mammalian Genotyping Service.

2.8 PCR and primer design

Polymerase chain reaction (PCR) uses thermostable *Taq* polymerase for the specific amplification of a precise defined DNA fragment between two short oligo-nucleotides which are called primer. The *Taq* Polymerase was discovered from the thermopile bacteria *Thermus aquaticus* [27, 158]. High optimum temperatures (75°C – 80°C) of *Taq* polymerase activity allow a cyclic repetition of

denaturation-, annealing-, and elongation-temperature and lead to an exponential amplification of selected DNA fragments. In this study, PCR reaction was set up with the AmpliTaq Gold® Polymerase, which is provided in an inactive state and becomes activated by an additional heating step at the beginning of the PCR (hot start). In order to prevent amplification of duplicated or homologous regions, primer designs were based on the Basic Local Alignment Search Tool (BLAST) results for each amplified SNP region. OligoAnalyzer1.1.2 and Oligo Explorer1.2 were used to support the general primer design based on the following recommendations:

- The primer melting temperature[2] (Tm) should be between 58-64°C. Sequencing primers were designed to lower Tm values at around 56°C but required to be at least 18 nt long to retain sufficient sequence specificity. For both PCR- and sequencing, the primer Tm was calculated based on the following simplified formula:

  $$Tm = 2°C \times (\text{number of } [A+T]) + 4°C \times (\text{number of } [G+C]) \quad (3)$$

- When possible, PCR-primer pairs were designed with similar Tm values and a GC content between 40-60%
- Complementarities of 2 or 3 bases at the 3'-ends of all primer pairs as well as complementary sequences within primers and between primer pairs have been avoided to reduce primer–dimer formation.

The Assay Design Software 1.0 was applied especially for Pyrosequencing primer design. The software predicts the formation of loops and dimers of the amplicon as well as unspecific binding sites of the Pyrosequencing primer on the amplicon. The sequencing primers were designed based on the following guidelines:

- Primer length was chosen between 15 – 20 bp and the Tm between 42°C and 50°C
- The sequencing primer was placed close to the SNP and was designed complementary to the biotin-labeled strand
- Orientation and the position of the sequencing primer were chosen in this way that the hereby produced pyrogram give a clear interpretation (see chapter 2.9).

2.8.1 General PCR set up and control of PCR product

All DNA amplification reactions (excluding the Y-filer multiplex kit) were set up with: 1× GeneAmp® PCR buffer (containing 1.5 mM $MgCl_2$), 0.25 mM dNTPs, 0.4 µM of each primer, 1U AmpliTaq Gold ® Polymerase and approximately 10 - 20 ng of genomic DNA. DNA fragments obtained from PCR prior

[2] Tm defined as the temperature where 50% of the primers are bound, 50% are not bound.

to sequencing were analyzed by horizontal 1.4% agarose gels. Agarose gels were prepared by melting the agarose in 1xTBE buffer in a microwave oven, supplementing with 10 mg/ml ethidium bromide and pouring it onto a horizontal gel bed with comb slots. Next, 2 µl of PCR product were mixed with 5 µl 1x loading buffer (3x loading buffer contains: 10% glycerol, 2,5% ficoll, 0.025% bromophenol blue, 0.025% xylene cyanol) and then loaded onto the gel. The 1 kb Hyperladder IV or the 500 bp Hyperladder V was used as a DNA size marker. The gels ran in 1xTBE at 80 – 140 V (1-10 V/cm) and the DNA fragments were visualized on an ultraviolet (UV) transilluminator.

Laboratory facilities temporary allowed the analysis of the PCR products with the QIAxcel instrument and the DNA Fast Analysis Kit. The QIAxcel uses a capillary elctrophoresis approach for the separation of the DNA fragments. Similar to traditional gel-electrophoresis an electrical field is applied to the capillaries and causes a migration and separation of the DNA fragments stained with ethidium bromide (0.1%) based on their size. A photomultiplier detector at the end of the capillaries then detects the passing of the migrating DNA fragments by their fluorescence. Afterwards the data is converted to an electropherogram and a virtual gel image is generated by the QIAxcel ScreenGel Software. Before running the sample the QIAxcel system has to normalize the signal intensities across all twelve channels of the gel cartridge. For calibration, 15 µl of QX Intensity Calibration Marker were loaded into a 0.2 ml twelve tube stripe, covered with a drop of mineral oil and placed into the buffer tray at position MARKER 2. Calibration run was then started and analyzed by the ScreenGel Software. For the sizing of the DNA fragments the system uses two markers: An alignment marker, which is injected before the PCR product for each sample and a sizing marker which is injected only once as a separate sample. Electrophoresis was carried out by adding 18 ml of QX separation buffer and two times QX Wash buffer into the buffer tray and covering them with 2 ml of mineral oil. Then 15 µl of QX Alignment Marker were loaded into a 0.2 ml twelve tube stripe, covered with mineral oil and placed in the buffer tray at the position MARKER 1. PCR samples and a DNA Size Marker 50 bp – 1.5 kb were placed into the sample tray and electrophoresis was started with the software.

2.8.2 PCR prior Pyrosequencing

PCR amplifications were carried out in a 25 µl reaction volume with 0.5 mM primer concentration and one primer labeled with biotin (Table 2-2). Thermocycling parameters were: 95°C for 10 min, followed by 8 cycles of: 95°C for 15 s, 60°C for 15 s, and 72°C for 30 s; then, 40 cycles of: 95°C for 15 s, 58°C for 15 s, and 72°C for 30 s; and finally, 70°C for 5 min. For the amplification of the rs33963329 locus, PCR was set up with 0,8 µM of primer. Thermocycling parameters for rs33963329 and P224

were changed to: 95°C for 10 min, followed by 40 cycles of: 95°C for 15 s, 54°C (rs33963329) or 58°C (P224) for 15 s, 72°C for 30 s; and finally, 70°C for 5 min (Appendix-Figure 7-1). PCR products were controlled for appropriate size on a 1.4% agarose Gel.

Table 2-2: List of primers used for PCR prior Pyrosequencing.

variation	chromosome position[1]	forward primer (5´- 3´)	reverse primer (5´- 3´)	PCR product (bp)	sequence primer	orientation of sequence primer
P29	13007215	AGATGATAGACAGGTGATAGAT	Bio-CCCAGATATTTGGTCAAACAC	519	CTCATCTAAAAATACCTT	forward
P30	13006761	AGATGATAGACAGGTGATAGAT	Bio-CCCAGATATTTGGTCAAACAC	519	GTGATAGATAAGTTGAT	forward
P38	12994387	Bio-TCCCTTCTTTGTAGAGGACAA	CCCAGGCTGGTCTCAAACA	144	GTGCTGGGATGACA	reverse
P40	12994402	Bio-TCCCTTCTTTGTAGAGGACAA	CCCAGGCTGGTCTCAAACA	144	CTTCCTTGGCCTCCC	reverse
P41.2=M359	13001679	GGACCAACGAGGCTGCAGG	Bio-GCCCAACTCCTCTTCCCATC	223	AAGCATAGTGATAGGG	forward
M170	13357186	TTGCAGCTCTTATTAAGTTATG	Bio-TAAACTTGCCAATTACTTTCAAC	167	TACTTAAAAATCATTGTTC	forward
M173	13535818	AAAGTTGATGCCACTTTTCAGA	Bio-GCAGTTTTCCCAGATCCTGA	242	AATATTAACAGATGACAAAG	reverse
M198	13540146	Bio-ATTCCAGTCATGATGAGGTGG	ATGCCGTTTGCCTAGGTTAG	138	ACTTAAATTAACCTAAAAGA	reverse
P224	15795387	Bio-TCTAATTTATTTTTCTGGTTACTC	TAGCATTCACAGAATAAGGTGAC	146	GTGGTTTCAGTCAGCAGGG	reverse
P240	13108816	CTTGGGTATCTGTCTAGAAGT	Bio-CAAAATTCAGACCGAGTATGTA	189	CTTTCAGATCAATAACGT	forward
P244	12943108	AGCATCTAAGGTGTGCCACG	Bio-TTCTTACTCGCCGCTCTATC	346	CAGTGCAACAGGACC	forward
M343	2947824	TCTGATTCGCACAAGGCTC	Bio -CACCTTTGTCCTCTTGCTC	194	TGCCCTCGTGTTCCA	forward
Tat	13431977	GACTCTGAGTGTAGACTTGTGA	Bio-GAAGGTGCCGTAAAAGTGTGAA	112	TGTAGACTTGTGA	forward
rs2040607	13039638	GTATCCGAGACTCAGACTTC	Bio-GTTGTTTCTTCAGGGCAAGGT	190	GCAAAGACAGACCAGC	forward
rs35753737	12579277 - 12579278	TCCTAAAAGTCTGTTTCTGCTG	Bio-GTGCAAACCCTACTGCTTCG	202	AGCCCATGAGTTTGAA	forward
rs34485380	13104633 - 13104634	Bio-CCACTAGTTCAGGAGTTATACG	AAGGTCAGGAGTTCGAGACTA	274	GGAGAATTGCTTGAACCT	reverse
rs33963329	15774238 - 15774239	CAAACAAAATCCATATGTAAGAC	Bio-TGTGCATATGTACCCTAAAACG	181	CAATGCAGTTACCTATTTCT	forward

(1) Chromosome position is given after NCBI database build 36, hg18.

2.8.3 PCR prior Sanger sequencing

PCR amplifications were carried out in a 10 µl reaction volume with 0.5 µM of unlabeled primers (Table 2-3). Samples for Sanger sequencing were prepared as follows: a denaturation step at 95°C for 10 min two cycle steps; then thermocycling: first, 5 cycles of: 96°C for 20 s, 65°C for 45 s, and 72°C for 3 min; second, 20 cycles of: 96°C for 20 s, 60°C for 50 s, and 72°C for 3 min; and finally, 9 cycles of: 96°C for 20 s, 55°C for 1 min and 72°C for 3 min. PCR prior to Sanger sequencing of the variation MID504 and rs13304202 was set up like general PCR (chapter 2.8.1) but with additional 1 mM $MgCl_2$. Thermocycling parameters were: 95°C for 10 min, followed by 8 cycles of: 95°C for 15 s, 60°C for 15 s, and 72°C for 30 s; then, 40 cycles of: 95°C for 15 s, 58°C for 15 s, and 72°C for 30 s; and finally, 70°C for 5 min. After amplification the PCR product was cleaned-up with ExoSAP-IT®. For this, 3 µl of ExoSAP-IT were added to the PCR products and incubated for 30 min at 37°C. The ExoSAP-IT® solution contains two hydrolytic enzymes: the Exonuclease I and the Shrimp Alkaline Phosphatase (SAP). Exonuclease I digests residual single-stranded primers and any extraneous single-stranded DNA produced in the PCR. SAP dephosphorylates remaining dNTPs from the PCR mixture into ddNTPs, thus

making them non-reactive in further enzymatic steps. Reactions were stopped through incubation for 15 min at 80°C.

Table 2-3: List of primers used for PCR prior Sanger sequencing.

sequenced segment	chromosome position[1]	forward primer (5´- 3´)	reverse primer (5´- 3´)	PCR Product (bp)	sequence primer forward (5'-3')	sequence primer reverse (5'-3')	chromosome specificity
Seq. S1 15-630	12972989-12973604	TATAAACAAACACATGCTTACTG	ACAATGAAGTCCAGGTGCAG	615	TGACTCCAATGTCCTCTC	TGCAGTTCCAGCACTTTG	Y
Seq. S2 418-1174	12973392-12974148	CTCAAGCAGTCCTCCCACCT	TGTTGATTCAGACACCTGGAG	777	CAAGTAGCTGGGACTACA	GAAATCTTTAGCCAACTGAA	Y
Seq. S3 958-1785	12973932-12974759	CAGACAAGAGAGTTCTACCTGT	CCTGGGCAACAGAATGAGACT	828	CTACCTGTTTGACTTTATCA	AAAGAAAAGGAAGAGAGACT	Y
Seq. S4 1677-3116	12974651-12976090	ACAGTCTCTTTTATTTGATGGCA	AGATCACTGGGCTCGGGACA	1459	AGTCTCTCTTCCTTTTCTTT	CAGGAGTAAGGTTTCAGAG	Y
Seq. S5 2924-3875	12975898-12976849	CTATTGTGCCTGCTGAGTGC	CACAGATGATAGATAGATTAGATA	952	GCATCCCCAAAATGATCCA	TAGATAGATAGATAGGTAGAT	Y
Seq. S6 3669-4801	12976643-12977775	CTGTCTATCCTTCCATGTATCT	TTGTGCAAGAGTCTTGGTCTGT	1133	ATAGTTTTCTATTTATCTGTGA	GCAACCTCCATCTCCCAG	Y
rs13304202[a]	12926071 - 12926335	ATATATGCACCATTGATATACC	CTTTATACCCTTGTCACATGC	265			Y
MID504[a]	3189981-3190297	AATACATTACATACACATTTCCC	GTTTATAGTAAAGTTTCCAAGTC	316			Y
	3189980-3190397	AAATACATTACATACACATTTCCA	CAGTATCACTATTTTCATGGCAT	417			X

(1) Chromosome position is given after NCBI database build 36, hg18. (a) PCR primers were also used as sequencing primer.

2.8.4 PCR amplification of DYS495, DYS504, DYS536, and the Y-filer multiplex kit

PCR amplifications for the markers DYS495, DYS504, and DYS536 were carried out in a 10 µl reaction volume with 6-FAM labeled forward primers (Table 2-4). Thermocycling parameters for DYS495 were: 95°C for 10 min, followed by 30 cycles of: 95°C for 30 s, 60°C for 45 s, and 72°C for 1 min and finally, 60°C for 55 min. In the case of DYS504 the following thermocycling parameters were used: 95°C for 10 min, followed by 30 cycles of: 95°C for 30 s, 59°C for 30 s, 72°C for 30 s; and finally, 60°C for 55 min. For DYS536, the thermocycling parameters were: 95°C for 10 min, followed by 30 cycles of: 95°C for 1 min, 54°C for 1 min, 72°C for 1 min; and finally, 60°C for 45 min. Multiplex PCR of the AmpFLSTR® Yfiler kit was performed in accordance with the manufacturer's instructions. Hereby, 7,5 µl of PCR puffer were mixed with 3,5µl primer mix and 7.5µl of template. During multiplex PCR 17 different fragments were amplified and labeled with one of the dyes: 6-FAM, VIC, NED or PET (= dye set G5). All amplified STR-systems were analyzed on the ABI Prism®3130-Avant Genetic Analyzer (see chapter 2.12).

Table 2-4: List of primer used for Y-STR typing.

marker	chromosome position[1]	forward primer (5´- 3´)	reverse primer (5´- 3´)	PCR product in (bp)
DYS495[159]	13520576 - 13520787	6FAM-CCCAGCTATTCAGGAGGTTG	GCCAGAAAGTGTGAGTCATCC	212
DYS504[160]	2963148 - 2963439	6FAM-TCTACACCACTGTGCCAAGC	GGCAACAGAGCAACCCTCT	292
DYS536[2]	2974846 - 2975122	6FAM-ATCGCATTCCCCTCTCCTAC	TTGCTTTTCTGCTTCCCTTC	277

(1) Chromosome position is given after NCBI database build 36, hg18. (2) Primer sequences have been obtained from the Human Organizes Whole genome Database.

2.9 SNP Analysis by Pyrosequencing

Pyrosequencing of DNA fragments uses a direct-primer extension approach which detects released Pyrophosphate (PPi) in real-time when nucleotides enzymatically being incorporated. The method was for the first time described by Ronaghi, Uhlén and Nyrén in 1998 which developed a four-enzyme mixture for the performance of Pyrosequencing without washing steps [161]. The detection of one nucleotide-species occurs by four principal reactions: PPi, which is released during the DNA polymerase-catalyzed reaction, is detected by ATP sulfurylase and luciferase in a coupled reaction. The amount of light produced in the luciferase-catalyzed reaction can be detected by a CCD (charge-coupled device) camera of the Pyrosequencing instrument. The apyrase enzyme then hydrolyzes any non-incorporated nucleotides and remaining ATP in order to prevent background signals (Figure 2-2A). One important step towards the Pyrosequencing technique was the substitution of dATPαS for dATP, since dATP is also a substrate for the luciferase enzyme and therefore produces nonspecific signal [162]. Furthermore, Pyrosequencing requires that templates are bound to a solid surface. For this purpose, the template is labeled with biotin during PCR and is then bound to streptavidin-coated magnetic beads. The method – based on the underlying chemistry - cannot distinguish between the incorporation of dATPαS, dCTP, dGTP, and dTTP. For this reason the four different nucleotides are added stepwise.

The Pyrosequencing results are collected by the PSQ 96MA2.1.1 software and presented in form of a pyrogram. Each signal peak of the pyrogram presents the incorporation of one nucleotide species which is labeled based on the dispersion order. The dispersion order describes the order in which the single nucleotides are added to the sequencing reaction. The height of each signal peak indicates the number of added nucleotides. For example, a repetition of the same base (double or triple bases) in the sequences should result in a two- or threefold higher peak (Figure 2-2 B).

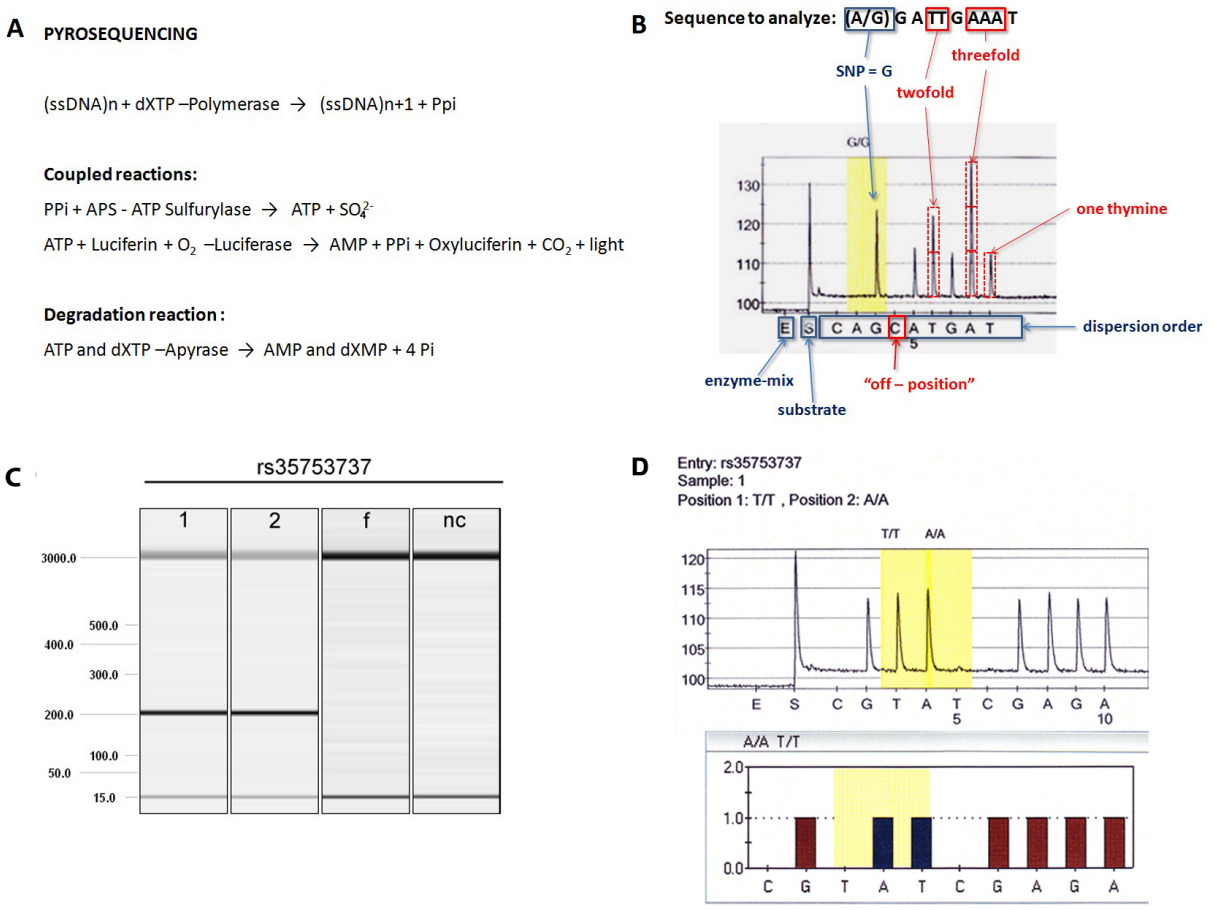


Figure 2-2: Principle of Pyrosequencing.
(A) The Pyrosequencing cascade reaction system. (B) Description of a pyrogram as it is obtained after Pyrosequencing and the analysis with the PSQ 96MA2.1.1 software. (C) Pyrosequencing at the example of the detection of the SNP rs35753737 (TA/AT). Gel-electrophoresis of the PCR reaction prior to Pyrosequencing with the QIAxcel instrument. The 15 bp and 3000 bp lines are the alignment markers which indicate that the electrophoresis is started and completed. Primer-specificity for the Y-chromosome was tested with male (number 1 and 2) and female DNA (f). (nc) is negative control. (D) The upper graphic presents the pyrograms for the variant TA, the lower graphic shows the histogram for the theoretical expected second allele variant AT, which was not observed.

Before the Pyrosequencing reaction, the target sequence was amplified (chapter 2.8.2) and a protocol file with the dispersion order was generated for each SNP entry with the PSQ 96MA2.1.1 software. Hereby, it was tried to avoid cases in which the determination of the SNP's identity depends only on the distinction of the signal intensity. Furthermore, the PSQ 96MA2.1.1 software sets an "off-position" after the SNP to avoid that the pyrogram is shifted out of phase (Figure 2-2 B). After the amplification of the target sequence (Figure 2-2 C), single stranded DNA was prepared from the PCR products and hybridized with sequencing primers with the help of Pyrosequencing Sample Prep Tool and magnetic sticks on a 96 well plate workstation. Briefly, 20–25 µl of PCR product was added to 33 µl of 2× binding buffer and 8 µl of streptavidin-covered magnetic beads. The DNA was bound to the beads after shaking the mixture for 15 min at 65°C. Next, the DNA was denatured by

transferring the beads to 0.5 M NaOH and incubated for 1 min. Afterwards the beads were washed for 5 min in 50 µl annealing buffer. The beads with single stranded DNA were then transferred to 50 µl of annealing buffer containing 2 µM sequencing primer, and the mixture was shaken for 4 min at 80°C. Pyrosequencing was carried out on the PSQ 96 MA robot. For this, the reaction cartridge was loaded with the enzyme mix, the substrate solution and the single nucleotides. The amount of the single solutions is charged by the dispersion order:

Amount of nucleotide in µl = number of nucleotides in probe x 0.2 + 50 µl (4)

Amount of enzyme/substrate in µl = number of samples x 5.5 + 50 µl (5)

The Pyrosequencing runs were performed based on the standard parameters: Enzyme pulse time = 97 ms, substrate pulse time = 88 ms, reagent pressure = 400 mbar, nucleotide pulse time = 8,5 ms, nucleotide pressure = 650 mbar, block temperature 28°C, mixer frequency = 35 Hz, cycle time 65 s, reagent priming pulse time = 100 ms, nucleotide priming pulse time = 30 ms. Pyrosequencing results were then analyzed with PSQ 96MA2.1.1 software (Figure 2-2D).

2.10 SNP Analysis by Sanger sequencing

Sanger sequencing was used to type the variations MID504, rs2757249, rs2534929, rs2534930, rs2534931, and rs5940712. The sequencing reaction was performed with the BigDye®Terminator (BDT) v1.1 or v.1.3 Cycle Sequencing Kit. The chemistry of the BDT Cycle Sequencing kit is based on the chain termination method which uses non-extendable ddNTPs as DNA chain terminators and was developed by Frederick Sanger [163]. Important enhancements of the Sanger sequencing method are the labeling of the ddNTPs with four different fluorescent dyes and the thermocycling by the application of a *Taq* polymerase. The BDT Kit uses a two component-dye-system which consists of a fluorescein donor dye, e.g. 6-FAM, linked to one of four dRhodamine acceptor dyes (Table 2-5).

Table 2-5: Dyes included in the BigDye sets F and E.

Terminator	Acceptor Dye
A	dichloro[R6G]
C	dichloro[ROX]
G	dichloro[R110]
T	dichloro[TAMRA]

The excitation maximum of each dye label is that of the fluorescein donor and the emission spectrum that of the dRhodamine acceptor [164]. Next to the dye terminators the ready reaction buffer contains dNTPs, magnesium chloride, r*Tth* pyrophosphatase and FS-*Taq* polymerase. FS-*Taq* polymerase shows a substitution of phenylalanine by a tyrosine at position 667 and incorporates chain-terminating didesoxyribonucleoside triphosphates (ddNTPs) much more efficiently than the wild-type *Taq* DNA polymerase [165]. The dNTPs mix of the BDT kit substitutes also dITPs in place of GTPs to minimize band compressions and dUTPs in place of TTPs to improve the incorporation of the T terminator. The reaction was carried out in a 10 µl volume with 1 µl of BDT Ready Reaction Premix, 1.5 µl BDT BigDye sequencing buffer, 3 µl of the ExoSAP-IT treated PCR product and 4.5 µl of water. The manufacturer's parameters for double-stranded DNA were used for thermocycling: 96°C for 1 min, 25 cycles of 96°C for 10 sec, 50°C for 5 s, and 60°C for 4 min. The pGEM®-3Zf(+) template DNA and the -21 M13 forward primer, which are included in the kit, were used for positive control. The sequencing product was purified by ethanol precipitation. Briefly, 2 µl of sodium acetate and 25 µl of absolute ethanol were added to the PCR product and centrifuged for 30 min at 2000 × g. Ethanol was removed by turning the tubes on filter paper and briefly centrifuging at 100 × g for 10 s. Precipitated DNA was washed with 50 µl of 70% ethanol for 5 min at 2000 × g. Again ethanol was removed by briefly centrifugation of turned tubes at 100 × g for 10 s. Samples were dried at room temperature for 15 min and resuspended in 15 µl of Hi-Di™Formamide. Sequencing analysis was conducted according to the manufacturer's settings on the ABI Prism®3130-Avant Genetic Analyzer (chapter 2.12).

2.11 Sequencing of the HLA loci

HLA loci were sequenced for the exons 2-4 of HLA-A and HLA-C, exons 2 and 3 of DQB1, and exon 2 of DRB1. Sequencing was carried out with the Allele SEQR kits which contain ready reaction mixes with specific primers for all HLA classes. The Allele SEQR kits are based on the principle of Sanger sequencing and use the BigDye chemistry (chapter 2.10). Prior to sequencing reactions the target HLA region was amplified by PCR. For this, 16 µl of the locus specific PCR Mix was mixed with 4 µl of DNA solution. For the HLA loci DQB1 and DRB1, only 8 µl of PCR mix were mixed with 2 µl of DNA, because fewer products were required (fewer exons have to be sequenced). HLA sequencing from HSE samples were carried out by using the entire sample volume for PCR. For this purpose the HSE sample was concentrated from 50 µl to the required amount by using the SpeedVac centrifuge.

HLA locus-specific PCR was carried out based on the following parameters: 95°C for 10 min, followed by 36 cycles of 96°C for 20 s, 60°C for 30 s, and 72°C for 3 min. The PCR products obtained were checked for their expected fragment size: HLA-A – 2 kb, HLA-C - 1.2 kb, DRB1 – 300 bp, and DQB1 – 300 bp by gel-electrophoresis (chapter 2.8.1). Prior to sequencing, the PCR products were cleaned up with ExoSAP-IT (chapter 2.8.3). Sequence reactions were set up in 0.2 ml PCR stripes with 8 µl of ready-prepared exon-specific sequencing mix and 2 µl of PCR product. Forward and reverse sequence reactions were carried out separately. For example the sequencing of the exons 2, 3, and 4 of HLA-A required the set up of six sequence reactions. Thermocycling was carried out by 25 cycles of 96°C for 20 s, 50°C for 30 s, and 60°C for 2 min. Sequencing products were precipitated and resolved in 20 µl Hi-Di-Formamid as already described in chapter 2.10. Sequence analysis was conducted according to the manufacturer's settings on the ABI Prism®3130-Avant Genetic Analyzer (chapter 2.12). The HLA alleles were determined based on the sequencing data after capillary electrophoresis with the program Assign ™ DBT 3.5.1.

2.12 Capillary electrophoresis

Capillary electrophoresis (CE) was carried out with the ABI Prism®3130-Avant Genetic Analyzer after manufacture's settings. Similar to traditional electrophoresis, an electrical field is applied to a gel, so that the negatively charged DNA fragments are separated by moving towards the positive electrode (anode). The speed at which the DNA fragments are moving are inversely proportional to its molecular weight. The first CE separations of DNA were performed in the late 1980s. Since the introduction of new CE instrumentation in the mid-1990s, the technique has gained rapidly in popularity for routine forensic analyses [50]. The advantage of CE to traditional agarose-gel electrophoresis is the improved separation of the DNA fragments by size with the maximum possible resolution of one base. In contrast to agarose-gel electrophoresis, CE uses thin capillaries (50 µm inner diameter) which are filled with a low viscosity polymer and connected to a high voltage power supply. During CE, the final detection of the DNA fragments is linked to the emission of fluorescence of the single fragments when they pass a detection window at the end of the capillary before the anode buffer reservoir.

All CEs therefore require the labeling of DNA fragments with a fluorophore / dye terminator, as described before (chapter 2.10). In the ABI Prism®3130-Avant the excitation of the fluorophore occurs by an Argon laser and the emission is detected by a CCD camera (Figure 2-3). Because each dye emits light at a different wave-length when excited by a laser, all four (Sanger sequencing) or five

(detection of STRs) colors can be injected and distinguished in a single capillary injection. The emitted light energy is converted to an electronic signal that is proportional to the amount of light emitted based on the number of dye-labeled molecules present. These signals are measured in relative fluorescence units (RFUs) and make up the peaks seen in capillary electropherograms. With a spectral calibration for each capillary a matrix is created that is used during a run to reduce raw data from the instrument for the 4- or 5-dye data stored in a sample file. A multicomponent spectral calibration is performed by testing a standard set of DNA fragments labeled with each individual dye, known as matrix standard samples. The data from each of the dyes creates a matrix or spectral calibration file that reflects the color overlap (bleed-through) between the various fluorescent dyes [50]. In this study, spectral calibration was carried out with the matrix dye sets DS33-G5 which contain the five dyes used in STR analysis and with the matrix E_BigDye_v1.1 and Z_BigDye_v1.3 which contain the dye sets E and F for Sanger sequencing (chapter 2.10 and Table 2-5). For calibration, 190 µl Hi-Di Formamide were mixed with 10 µl Dye set Matrix standard and incubated for 5 min at 95°C. Then the appropriate dye set was loaded 16 times (for the 16x capillaries array) in a 96 septa well, closed with a clean plate cover and briefly spun down. Next, the Septa well was placed in the Plate Base and Retainer and calibration run was started by the ABI Prism Data Collection Software. The electrophoresis protocol of the samples depended on the type of sample and analysis. For Sanger sequencing samples were cleaned by precipitation and resolved in 15 µl Hi-Di Formamide (chapter 2.10). The solved samples were directly loaded into the ABI-plate systems and run after sequencing protocol. For the analysis of multiplex PCR for STR profiling it is necessary to determine exact fragment length of the PCR products. Because electrophoresis is a relative rather than an absolute measurement technique each sample is mixed with a constant set of DNA fragments of known size. This internal size standard is colored with an extra dye and allows the correlation between detection time and fragment length. Furthermore CE of STR analysis includes the run of a kit specific allelic ladder, which is a mixture of common alleles for all typed STR loci. The use of an allelic ladder allows an automated assigning of the determined fragment length to certain STR alleles. In this study CE with STR samples were carried out by mixing the standard LIZ 500 (for AmpFLSTR® Yfiler multiplex analysis) or LIZ 600 (for DYS495-, DYS504-, and DYS536 singleplex PCR) at a ratio of 1:25 with Hi-Di Formamide. Next, 20 µl of the diluted standard and 1,5 µl of sample or allelic ladder were loaded in a 96 septa well and were ready for electrophoresis run after the manufacture protocol. Unused positions within two connected rows of the septa well were filled with plain Hi-Di Formamide to prevent the formation of air-bubbles in the capillaries. All electrophoresis runs were carried out with 36 cm capillaries filled with the denaturing polymer POP-4. CE was run with the

standard protocol HIDFragmentAnalysis_36_POP4 and 10 s of injection time. Data were collected with ABI Prism Data Collection Software 3.1.

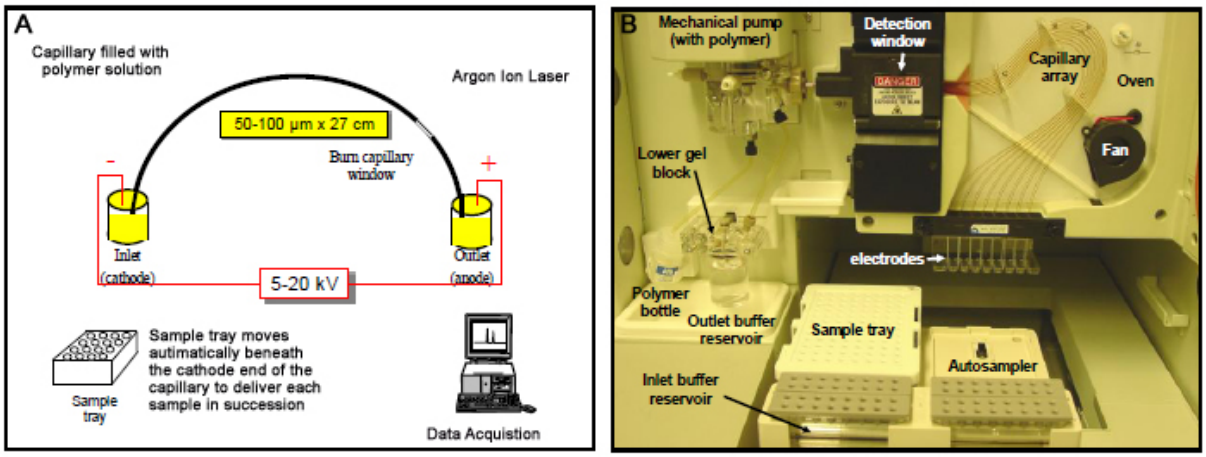


Figure 2-3: Schematic figure of capillary electrophoresis.
(A) The capillary is a narrow glass tube, filled with a viscous polymer solution that creates a sieving environment for DNA molecules. Samples are placed into a tray and injected onto the capillary by applying a voltage to each sample. A high voltage is applied across the capillary after the injection in order to separate the DNA fragments in a matter of minutes. The dye-labeled products are analyzed as they pass by the detection window and are excited by a laser beam. Computerized data acquisition enables rapid analysis as a digital storage of separation results. (B) Components of the CE in the ABI 3130cl Genetic Analyzer with a16-capillary array. Pictures are redrawn from Butler 2009 [50].

Sequencing data were processed with ABI Prism®Sequencing Analysis Software 5.2. Additionally, the software Assign™ DBT 3.5.1 was used for HLA typing. The sequencing software processed the fluorescence signals and assigned a base to each peak (A in green, C in blue, G in black, T in red, or N for base not determined). The data obtained during CE of STR samples were analyzed with the GeneMapper ID-x1.1.1 software, which uses the size standard to determine fragment lengths of the single PCR fragments. Detected light signals of the fragments were presented in an electropherogram as differently colored peaks. Hereby all peaks are colored based on the emitted light of the labeled fragments. Fragments which are labeled with 6-FAM emit blue, with VIC green and with PET red. Fragments labeled with NED emit yellow light but are presented as black peaks in the electropherogram. Peak heights and areas represent the intensity of the fluorescence signal. The analysis of the amplified fragments after AmpFLSTR® Y-filer PCR includes also an allelic ladder as well as defined bin-sets for the automatic determination of single alleles.

2.13 Probe design for haplotype-specific extraction

Allele-specific HSE probes were designed for a total of 13 Y-chromosomal SNPs and 5 variations located inside the HLA locus. Since the polymerase catalyzes the addition of nucleotides to the 3'-OH of completely annealed probes, all HSE probes were designed either in a forward or reverse

orientation with the discriminating nucleotide located at either the 3' (last) or penultimate (-1) position.

Table 2-6: All HSE probes used in this study.

A: Extraction sites on the Y-chromosome

probe[1]	mismatch	sequence	tested different probe-length in nucleotides[2]
P30FG	G-T	CAGGTGATAGATAAGTTGATCG	17, 19 - 22
P30FG -1	G-T	GGTGATAGATAAGTTGATCGA	21
P30FA	A-C	CAGGTGATAGATAAGTTGATCA	17, 19, 22
P30FA -1	A-C	AGGTGATAGATAAGTTGATCAA	21, 22
P30RC	C-A	TCTATCCATCTATCATCTATTTATC	19 , 21, 24, 25
P30RC -1	C-A	TCTATCCATCTATCATCTATTTATCG	23, 26
P38FA	A-G	GCTGGGAGGGTGGCTCCCGCA	10 - 13, 15, 16, 19 - 21
P38FA -1	A-G	GGGAGGGTGGCTCCCGCAT	12, 16, 19
P38FC	C-T	CTGGGAGGGTGGCTCCCGCC	10, 12, 13, 15, 18 - 20
P38FC -1	C-T	GGAGGGTGGCTCCCGCCT	12, 18
M170FC	C-T	TACTTAAAAATCATTGTTCC	20
M170FA	A-G	TACTTAAAAATCATTGTTCA	21
M173FA	A-G	TCAAGGGCATTTAGAACA	18
M173FC	C-T	TCAAGGGCATTTAGAACC	17
M173RT	T-C	CTGAATATTAACAGATGACAAAGT	24
M173RG	G-A	CTGAATATTAACAGATGACAAAGG	24
M198FC	C-A	AATTAATATTTTTGAAAGAGC	21
M198FT	T-G	AATTAATATTTTTGAAAGAGT	22
M198RG	G-T	CTGTACTTAAATTAACCTAAAAGAG	25
M198RA	A-C	CTGTACTTAAATTAACCTAAAAGAA	25
P224FC	C-A	TCAGAAATGAGTGTGACATCTTC	9 - 21, 23
P224FT	T-G	TCAGAAATGAGTGTGACATCTTT	18 - 21, 23
P224RG	G-T	GTGGTTTCAGTCAGCAGGGG	17, 20
P224RA	A-C	GTGGTTTCAGTCAGCAGGGA	17, 20
P240FC	C-A	TCTTTCAGATCAATAACGTCTC	17, 19 - 22
P240FC -1	C-A	TTTCAGATCAATAACGTCTCG	18 - 21
P240FT	T-G	TCTTTCAGATCAATAACGTCTT	22
P240FT -1	T-G	TTTCAGATCAATAACGTCTTG	20, 21
P240RG	G-T	GTAGGCTCAGATAAAGAACG	16 - 20
P240RG -1	G-T	TAGGCTCAGATAAAGAACGA	19, 20
P240RA	A-C	GGTAGGCTCAGATAAAGAACA	17 - 21
P244FA	A-C	CAGTGCAACAGGACCA	14, 16, 23
P244FA -1	A-C	AGTGCAACAGGACCAG	15, 16
P244FG	G-T	GCCCAGCAGTGCAACAGGACCG	11 - 16, 23
P244FG -1	G-T	GTGCAACAGGACCGG	13 - 15
P244RC	C-A	TATTGTCCTGCAGCTCCATCCCC	13 - 15, 23
P244RC -1	C-A	CAGCTCCATCCCCG	13, 14
P244RT	T-G	TTGTCCTGCAGCTCCATCCCT	11, 13, 16, 22
P244RT -1	T-G	CAGCTCCATCCCTG	14
M343FC	C-T	AGAGTGCCCTCGTGTTCCAC	20
M343FA	A-G	AGAGTGCCCTCGTGTTCCAA	20
TatFT	T-G	GTGTAGACTTGTGAATTCAT	20
TatFC	C-A	GTGTAGACTTGTGAATTCAC	18, 20
rs13304202FA	A-C	TAAGGAACATTACTCAAGAGA	18 - 21
rs13304202FA -1	A-C	TAAGGAACATTACTCAAGAGAC	20, 22
rs13304202FG	G-T	TAAGGAACATTACTCAAGAGG	21
rs13304202FG -1	G-T	GGAACATTACTCAAGAGGC	19
rs13304202RC	C-A	AGTTTTATTTATGGAGGAAGC	18, 19, 21
rs13304202RC -1	C-A	TTAGTTTTATTTATGGAGGAAGCC	19- 21, 24
S4:2701 RG	G-T	AGGCTGTGCTATTGATGAAAATG	20, 23
S4:2701 RA	A-C	AGGCTGTGCTATTGATGAAAATA	20, 23
S6:4204 FG	G-G	GGTCTTCCTCTGTTCCTCAG	17, 20
S6:4204 FC	C-C	GGTCTTCCTCTGTTCCTCAC	17, 20

B: Extraction sites at the HLA - locus

probe[1]	mismatch	sequence	tested different probe-length in nucleotides[2]
A_RA-1	A-C	TGTGTCCACTGTTCCAC	17
A_RG-1	G-T	TGTGTCCACTGTTCCGC	17
C_FAT-1	AT-GT	TCTTTCTGCAAAGGCATC	18
C_RAC-1	AC-GT	GCAGACACATTCAGACG	17
DR_RAA	AA-GG	ACTCACACCTTAGAACAAA	19
DR_RCC-1	CC-TT	CTCACACCTTAGAACCCA	19
DQ_RTG-1	TG-GT	GAGAAAGGTCTCATTACATGA	21
DQ_FTG-1	GT-TG	AGAGAGTGGCTGTTTGT	17
DQB1:05-06	T..T-C..C	GGCGACGACGCTCACCTCTCCT	16, 18, 22
DQB1:02-04	G..G-A..A	GGCGACGACGCTCACCTCGCCG	14, 18, 22

[1]Probes were named after the corresponding SNP, orientation (F = forward, R = reverse) and discriminating 3'-terminal base, which is also underlined in the sequence. Probe names have been denoted with -1 when the discriminating base is placed on the second base-pair of the 3'-terminus. [2]Numbers refer to the different lengths of the probes that were tested. For example, probe rs13304202FA was tested using lengths of 18, 19, 20, and 21 nucleotides, whereas the probes have been shortened step-wise on their 5'-termini.

For the investigation of the optimal probe design, several probe variations were designed for different probe lengths, which together were referred to one “probe set”. Different probe lengths in one probe set were tested by HSE in order to identify the probe length with the best separation effect. Table 2-6 provides the probe nomenclature, mismatch type, nucleotide sequence and sequence lengths used in this study.

For the prediction of best probe length, a melting curve analysis was performed for each specific probe or probe set in Visual OMP™ [38, 47]. For the simulation of melting curves (Figure 2-4C), this program calculates the concentration of hybridized probe (target-probe-duplex) with rising temperature based on the equations (6 through 8) as taken from SantaLucia, 2007 [47]. Duplex concentrations are calculated using the equilibrium equation. First, the equilibrium constant for the hybridization of probe to target can be calculated as follows:

$$K = \frac{[A + B]}{([Atot] - [AB] \times ([Btot] - [AB])} \quad (6)$$

where A and B indicate the two strands A and B in the random coil state, AB indicated the ordered (fully hybridized) duplex state, and Atot, Btot denotes the total concentrations of A and B.
Next, the equilibrium constant K can be computed from the Gibbs free energy ($\Delta G°_T$), also referred to free enthalpy.

$$\Delta G°_T = - RT \times \ln(K) \quad (7)$$

where R is the gas constant and T is temperature in Kelvin. The naught symbol (°) is added to indicate that the given energy values are for the idealized “standard state”, which means that the energy change refers to the amount of energy that would be released when one mole of substance reacts completely under standard conditions of 298K and a pressure of 1 atm. Next, $\Delta G°_T$ can be computed for any temperature T, if the enthalpy (ΔH°) and the entropy (ΔS°) are known based on equation (8):

$$\Delta G°_T = \frac{\Delta H° \times 1000 - T \times \Delta S°}{1000} \quad (8)$$

Values for ΔH° and ΔS° are predicted from the strand sequences involved in the duplex by applying the nearest-neighbor (NN) model. In addition to NN values for Watson-Crick base paired duplexes, Visual Omp™ uses empirical equations that allow the NN model to extend to salt concentrations, terminal dangling ends and all possible internal and terminal mismatches [45-48]. Based on this model, the simulation of HSE probe hybridization was performed for the hybridization temperature

and the conditions of HB buffer (Figure 2-4 B). Additionally, the starting concentrations of the probe and the template as well as sequence details have been entered in the program (Figure 2-4 A). Concentrations of target (con_{target}) and the molar mass (M) have been calculated using the following equations:

$$con_{target}\ (\mu M) = \frac{con_{DNA\text{-}Mix}}{M} \tag{9}$$

and

$$M_{template} = (\text{length of template in bp} \times 660\ \text{Da/bp}) \times 10^{-6}\ \mu\text{g/pmol} \tag{10}$$

where 10^6 Da presents 1 µg/pmol. Because in this study the target concentration con_{target} during HSE has to be calculated for the human genome, the template length is 3.2 x10^9 bp. With an input of approximately 300 ng genomic DNA in HSE, $con_{DNA\text{-}Mix}$ ≈ 10 ng/µl, the concentration of the human genome in HSE can be calculated after (9) and (11):

$$con_{human\ genome} = \frac{\approx 0{,}01\,\text{g/l}}{3{,}2 \times 10^9\ \text{bp} \times 660\ \text{Da/bp}} \approx 5\ \text{fM} \quad \text{haploid genome} \tag{11}$$

Furthermore to the simulation of the correct hybridization of the probe to the target (two-state model), Visual Omp™ is also able to calculate the concentrations of all occurring duplex formations in a multi-state coupled equilibrium calculation [47]. Thus, the change in duplex concentration for match and mismatch probes ($\Delta con_{M\text{-}MM}$) for a specific allele could be determined, as shown in Figure 2-4 C. In this study the target concentration for simulation of probe hybridization by Visual Omp™ was set up to be 3 fM, because the human template consists of a male DNA mix with two contributors. In addition to the calculation of the final match- and mismatch-duplex concentrations of HSE-probes, Visual Omp™ was also used for the determination of oligo-standard parameters:

- GC-content in %,
- $\Delta G°_T$ in cal/mol and
- Tm after the NN-model:

$$T_M = \frac{\Delta H°}{(\Delta S + R \ln C_T)} \tag{12}$$

whereas C_T is $C_A - C_B/2$ with C_A and C_B being the more and less concentrated strands, respectively [46].

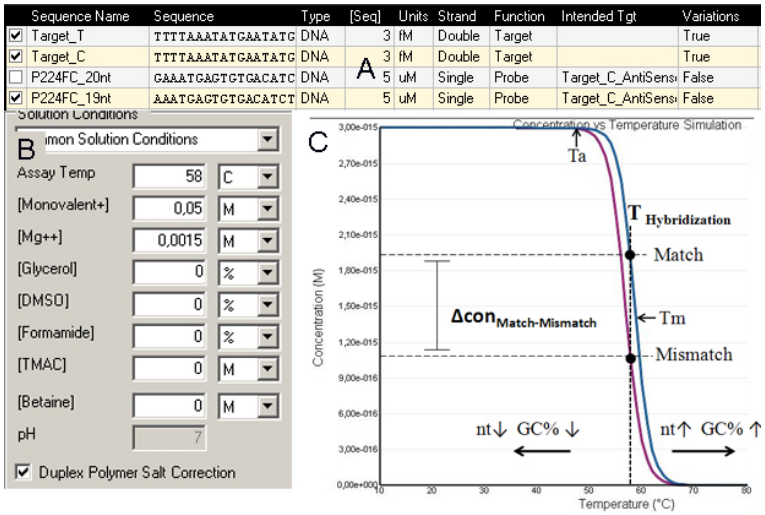


Figure 2-4: Direct comparison of the thermodynamic stability of match and mismatch using Visual Omp™.
The program allows the simulation of the concentration of hybridized matches and mismatches by adding the starting concentration (A) of the target and probe, as well as assay parameter (B) under the solution conditions. (C) Simulation of concentration versus temperature. Schematic representation of the comparison between match and mismatch by concentrations, Ta = optimal annealing temperature for PCR, Tm = melting temperature. Horizontal arrows indicate a switch in the hybridization curve for decreasing (↓) or increasing (↑) probe length (nt) and G/C content (GC%).

2.14 Haplotype-specific extraction (HSE)

DNA mixtures were aimed to be separated by the EZ1 HaploPrep Kit and a protocol designed by Dapprich *et al.* 2008 [166]. All standard extractions were set up in 30 µl reaction volume in 1,5 ml tubes with screw cap containing 200-400 ng genomic DNA, 5 µM of probe and 1x hybridization buffer (HB). In this study HSE reaction in general were set up with 1:1 mixed male DNA samples. Individual DNA samples were denatured at 95°C for 7.5 min on an external heating block True Temp with a heated lid. The BioRobot® EZ1 workstation was prepared for HSE extraction by loading the HSE protocol from the HaploPrep card (chapter 2.14.2) and inserting the EZ1 HaploPrep consumables: cartridges, elution tubes and tips with tips holders (Figure 2-5). After denaturation the HSE samples were transferred onto the heating block position of the EZ1 robot and incubated at 58°C for 20 minutes for allele-specific enzymatic extension and incorporation of biotin-dUTP. Next, samples were incubated with streptavidin-coated magnetic microparticles to capture the biotinylated amplicons. After incubation, the beads were washed twice with washing buffer and re-suspended in 50 µl elution buffer. At this point the samples were removed from the robot. For further analysis of the HSE samples the biotinylated amplicons were detached from the magnetic beads by incubating the

samples at 80°C for 10 min. After incubation, the HSE samples were put in a magnetic separator and the supernatant containing the extracted DNA was transferred without beads into a fresh 1.5 ml tube with a screw cap. Multiplex HSE reactions were carried out based on the HSE standard protocol but with the use of more than one probe in one reaction (Table 2-7).

Table 2-7: Standard HSE.

components	stock	final concentrations	input in HSE in µl
HB	2x	1x	15
probe	100 µM	5 µM	1.5
each additional probe	100 µM	5 µM	1.5
DNA mix	40 ng/µl	300 ng	7.5
Aqua as iniectabilia	-	-	add to 30 µl

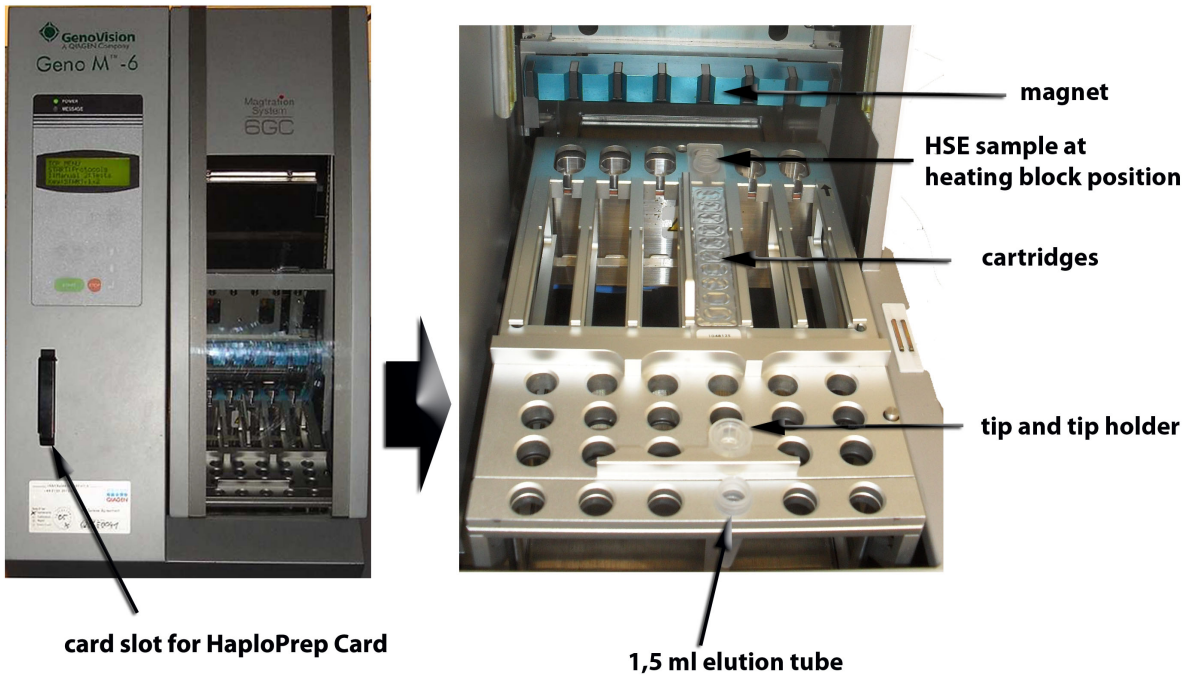


Figure 2-5: Preparation of the Biorobot® EZ1 (before Geno M™-6 from GenoVision) for the HaploPrep protocol.

2.14.1 Test of different hybridizationbuffers

HSE was tested first with different dilutions of the standard HB provided by Qiagen. Second, HB was also tested with varying concentrations of only one single component. HB dilutions of ½, ¼, and ⅛ were set up by using only 7.5 µl, 3.75 µl, and 1.88 µl of the Qiagen standard HB buffer compared to the standard amount of 15 µl, whereas DNA and probe concentration remained constant and water was added to a reaction volume of 30 µl. Different compositions of HB were tested by setting up

individual components of the buffer to various ratios related to the respective concentrations of the 1x standard HB: *Taq* polymerase: 8, 3, ½, ¼ , ⅛-fold; biotin-dUTP: 10, 5, 3, 2.2, 1.5, 0.75, 0.35-fold; and dNTPs: ½, ¼, ⅛-fold. The detailed buffer composition cannot be given in this work because of confidentiality agreement restriction.

2.14.2 Test of additional wash steps

During the HSE standard protocol, which is provided on the HaploPrep Card, the biotin-labeled DNA strands are bound to magnetic streptavidin beads. Beads with DNA are then robotically washed twice in 200 µl and 500 µl wash solution. Finally, the washed beads are resuspended in 50 µl elution solution. All steps are carried out automatically by the BioRobot EZ1 (Figure 2-6). In order to evaluate the possible effect of additional wash steps carried out by the EZ1, the instrument instructions in the protocol file were modified by adding commands for two additional wash steps and uploaded through an external PC via an RS232 interface.

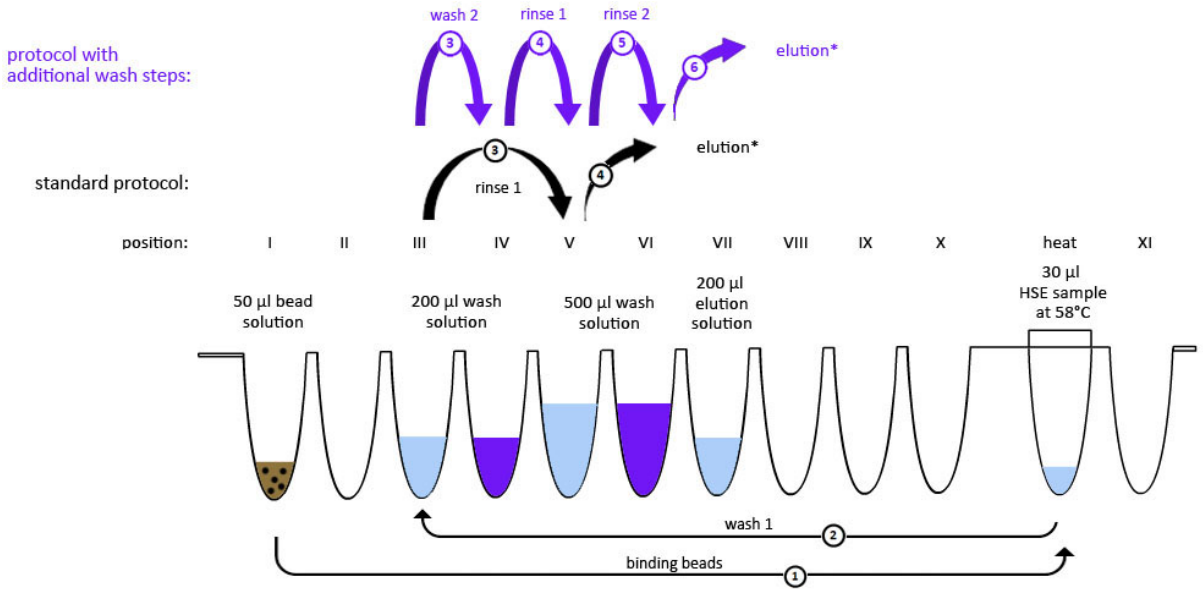


Figure 2-6: Comparison between HSE-EZ1-protocols with 2 and 4 wash-steps.
The figure shows a schematic drawing of an EZ1 cartridge. Available positions in the cartridge are labeled with roman numerals. Position one contains the beads solution, position three to six wash solution, and position seven the elution solution. The heating block for the incubation of the HSE sample at the hybridization temperature is located between positions ten and eleven. Arabic numbers indicate single HSE protocol steps which are carried out by the EZ1. Filled tubes in light blue as well as the corresponding black numbers and arrows indicate the steps performed by the standard protocol of the HaploPrep EZ1 kit. Filled tubes, arrows and numbers in purple describe the modified protocol with two additional wash steps. (*) indicates the step before the protocol start; the EZ1 robot pipettes 50 µl of the elution solution from position seven into position nine for the resuspension of the final HSE sample.

2.14.3 Analysis of HSE efficiency

Analogously to autosomal profiling, the identification of male individuals by Y-chromosomal haplotyping is based on the analysis of STR markers. Therefore, the separation success of male DNA mixtures was analyzed by the determination of Y-chromosomal STR markers after HSE. For first HSE testing the three markers, DYS495, DYS504, and DYS536, were used as extracted markers and analyzed in a manually set up singleplex PCR (chapter 2.8.4). These three additional markers were used for HSE analysis because of their close distance to the HSE probes, M343, M178 and M198, which were designed as first extraction probes. However, the used PCR protocol for the amplification of DYS495, DYS504 and DYS536 did not appear as very robust and sensitive. Therefore, further HSE probes were designed in the near of STR markers, which are included in the AmpFLSTR® Yfiler kit (Y-filer) from Applied Biosystems. The advantage of the use of the Y-filer kit is that it is one of the standard application kits in forensic investigations of males and presents a sensitive multiplex PCR system. In total the Y-filer kit includes the detection of 17 loci, which are distributed over the short and long arm of the Y-chromosome beyond of the heterochromatin regions. The distances between the STRs range from maximal 3.6 Mb to minimal 276 kb, except for the four markers DYS635, DYS437, DYS439 and DYS389I+II. These five markers are located very close to each other and show distances between 48 kb and 97 kb (Figure 2-7). Comprehensive validation studies by the BCA Forensic Science Laboratory and Applied Biosystems evaluate the AmpFLSTR® Yfiler multiplex kit as a robust and reliable system and analyzed the kit-specifications after sensitivity, stutter occurrence, balance and mixture analysis [70, 106, 167][3]. These published Y-filer specifications could be observed also in own validation studies, which were carried out in order to evaluate the Y-filer kit as a standard protocol for forensic investigation in our laboratory.

Sensitivity: Informative partial profiles can be obtained with amounts of >60 pg male DNA, whereas full profile detection was found for > 125pg DNA.

Stutter: The stutter values were determined specific for each loci and vary from 4.2% for DYS438 till 15.8% for DYS398II.

Balance: The amplification of each single locus is not of equal magnitude and shows some imbalance. Especially the marker DYS19, DYS392, DYS439 and DYS635 are reported to be less sensitive and are the first to be affected by reduction or drop out.

Male-Male mixture: The minor component of a DNA mixture containing two different male contributors can be detected up to a ratio of 1:10 till 1:20. Hereby, the ratio is mainly limited by the

[3] The here described Y-filer specification are not complete but present a selection after their significance in this study.

narrow concentration range in which the Y-filer kit shows a balanced amplification for all primer sets. Therefore, the detection limit depends on used concentration of major and minor component. For example, when the DNA input of the major component fit the optimal amount of 1 ng, than a ratio of 1:20 would allow only 50 pg DNA input for the minor component. Therefore the minor component reaches the limit of sensitivity of the kit. Furthermore, if the DNA input of major component would be decreased, the maximum ratio for the minor component would decrease also under 1:20. On the other hand, the limiting ratio of 1:20 cannot be expanded by increasing the concentration of the major component. Y-filer reactions with more than 1 ng DNA input start to show an unbalanced amplification.

In order to set up suitable male DNA mixtures for HSE which contain informative allele combination, first the Y-STR haplotypes were determined from laboratory standard samples (Table 2-8).

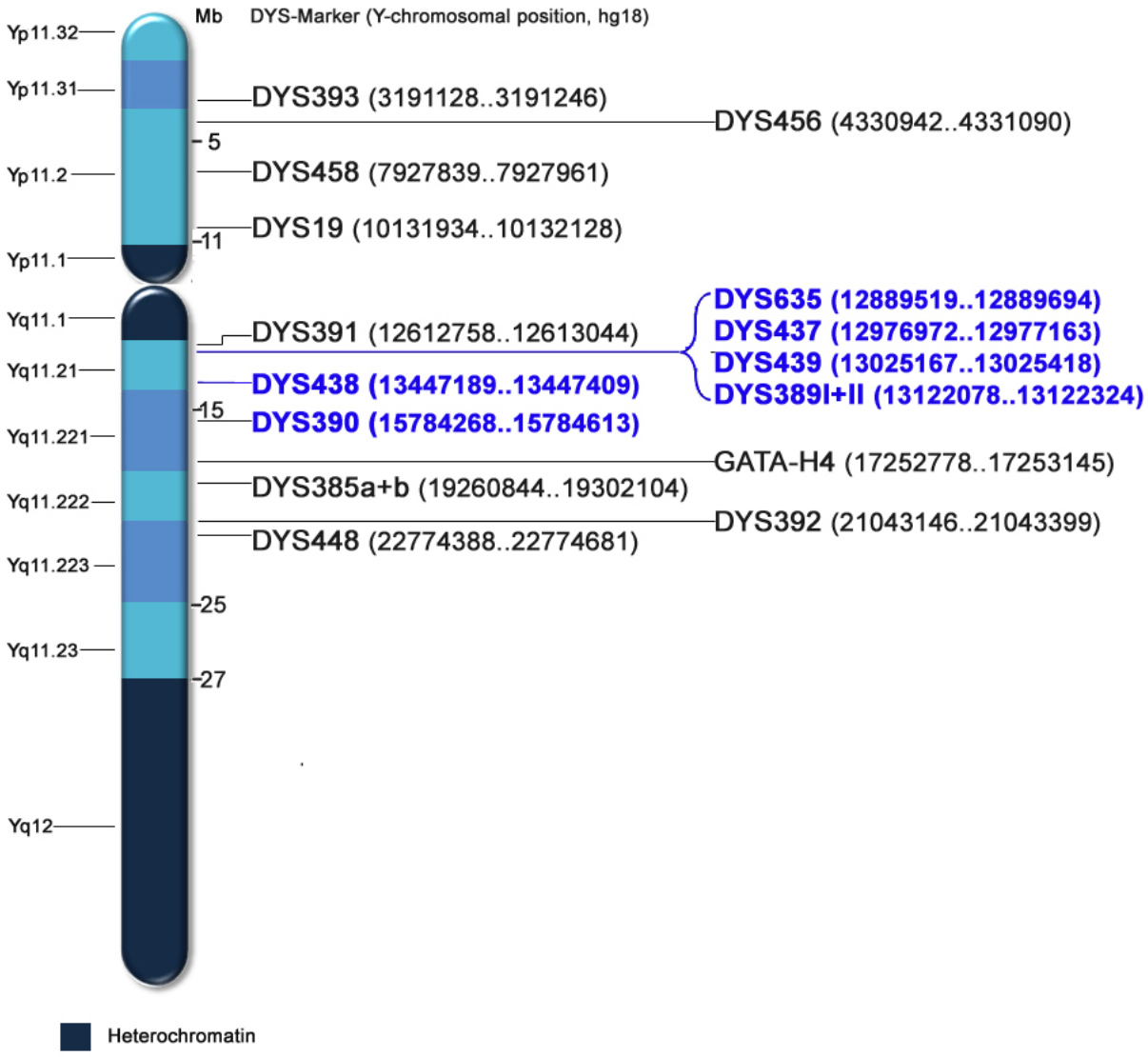


Figure 2-7: Map of Y-filer STR markers. In this study extracted markers are indicated in blue. Markers in black refer to non-extracted loci.

Table 2-8: Summary of Y-chromosomal haplotypes used in this study for the preparation of male DNA mixtures.

sample ID	ancestry	Y-filer																DYS495		DYS504	DYS536	
		DYS 456	DYS 389I	DYS 390	DYS 389II	DYS 458	DYS 19	DYS 385	DYS 393	DYS 391	DYS 439	DYS 635	DYS 392	GATA H4	DYS 437	DYS 438	DYS 448	size (bp)	repeats	size (bp)	size (bp)	repeats
1114	Germany	14	12	23	28	15	15	15,15	13	10	11	22	11	10	16	10	20					
1118	Germany	16	13	24	29	17	14	11,14	13	11	11	23	13	13	15	12	18	212	15	281	269	8
1573	Germany	14	12	22	28	16	14	13,14	13	11	11	21	11	12	15	10	21					
5572	India		13	24	29		15	12,16	12	10	11		11			9					273	9
5573	Germany	14	13	23	29	15	14	15,15	13	10	12	22	11	11	16	10	20	215	16	285	269	8
5574	Germany	16	13	25	31	14	17	11,14	13	11	10	23	11	13	14	11	20	209	14	273	269	8
94447	Germany	15	13	24	29	14	14	11,14	13	10	12	23	13	12	15	12	19					
95210	Scottland	17	13	24	30	17	14	11,14	13	10	13	25	13	13	15	12	19					
99309	Germany	15	14	24	30	17	14	11,14	13	11	12	24	13	12	16	12	19					
104122	Finnland	14	13	23	29	16	14	11,13	14	11	10	21	14	12	14	10	19					
110776	Germany	16	13	22	29	19	14	11,14	13	11	12	24	13	12	15	12	19					
116607	Italy	15	12	22	29	18	15,17	14	13	11	11	21	11	12	16	10	21					

The assignment of the number of repeats for DYS495 and DYS536 has been established according to gene bank references in the NCBI and STR databases. The fragment length obtained for DYS504 was not in accordance with the reference and therefore no numbers of repeats were determined.

2.14.4 Calculation of probe efficiency

The evaluation of HSE success derives from the analysis of the ratios between obtained peak areas of extracted versus non-extracted allele (or targeted versus non-targeted allele). Therefore the efficiency of allele-specific HSE probes was evaluated as a ratio per equation 13 or as a percentage of enrichment using equation 14.

$$\text{HSE success}_{\text{RATIO}} = \frac{\text{peak area of intended DYS allele}}{\text{peak area of unintended DYS allele}} \qquad (13)$$

In results the ratio is shown as the multiple of the extracted contributor allele, e.g. a ratio of 1:3 is illustrated as 3x (the extracted allele is 3 times of the non-extracted allele).

$$\text{HSE success}_{\text{ENRICHMENT}}$$

$$= \frac{\text{peak area of intended DYS allele}}{(\text{peak area of intended DYS allele} + \text{peak area of unintended DYS allele})} \text{ x100\%} \qquad (14)$$

The HSE success was then categorized after the enrichment as follows:

Complete separation: A separation was considered “complete” when HSE enrichment was >89% per equation 14. In this case the non-enriched contributor falls under a stutter threshold of 12%, which accord to the average of the published stutter values for the multiplex Y-filer kit [167]. Stutter

percentages are calculated by dividing the peak area of the stutter peak by the peak area of the true allele peak. In the case of a complete separation with an absence (or "dropout") of the non-enriched allele, the absence allele peak area was set to 200 for the calculation of HSE efficiency. In laboratory praxis a peak area of 200 was evaluated as a standard cut off value for background signals.

Significant separation: HSE success was classified as a "significant separation" when HSE enrichment was observed as ≥ 61% and enrichment of each individual contributor allele could be observed in parallel HSE reactions. The limit for a "significant enrichment" was set up based on observed allele balances within one Y-STR system after multiplex Y-filer analyses of 1:1 male DNA mixtures or HSE reaction. Hereby the imbalances between the two contributor alleles within one Y-STR marker were calculated for all Y-filer markers as HSE enrichment according to equation 14. Y-filer reactions were carried out from set up male DNA mixtures. Perfect allele balance would be expected at 50%, which would mean equal peak size for both contributors. However mixture analysis showed that the deviation lied between 55 and 66%. Hereby the in this study targeted marker showed a maximum derivation of 61%.

Table 2-9: Variation of allele amplification of 1:1 male DNA mixtures.

		DYS 456	DYS 390	DYS 389I+II	DYS 458	DYS 19	DYS 385a+b	DYS 393	DYS 391	DYS 439	DYS 635	DYS 392	GATA H4	DYS 437	DYS 438	DYS 448
no HSE	n	59	59	20	59	12	45	2	51	49	60	49	57	57	59	48
	mean	57 ±5	56 ±5	61 ±6	59 ±4	66 ±4	55 ±4	53 ±1	56 ±5	56 ±4	55 ±4	55 ±3	56 ±5	55 ±4	57 ±5	55 ±4

In this study extracted markers are indicated in blue.

No separation: HSE success was classified as "no separation" when enrichment of one contributor was < 61%.

2.14.5 Calculation of background signals

In this study, seven of the 17 Y-filer loci (DYS635, DYS437, DYS439, DYS389I+II, DYS390, DYS438) were aimed for separation by HSE (Figure 2-7). Depending on the used extraction probe, these seven loci are described as extracted markers or targeted markers, because specific extraction probes were located in close distance. Hereby the term extraction site refers to the binding site of the probe and to the position of the Y-SNP (Figure 2-8). The distance between tested extraction sites and targeted markers varied between 150 bases and less than 100 kb. Nagy *et al.* 2007 [3] mentioned observed separation for markers in more than 40 kb distance and Dapprich *et al.*2008 [166] describe a drop down of the separation effect for distances > 30 kb. Therefore 100 kb were selected as the maximum extracted fragment length and all markers outside of 100 kb are here described as non extracted or

non targeted markers. Any signal detection of these non-extracted markers by the Y-filer multiplex PCR after HSE referred to HSE background signals.

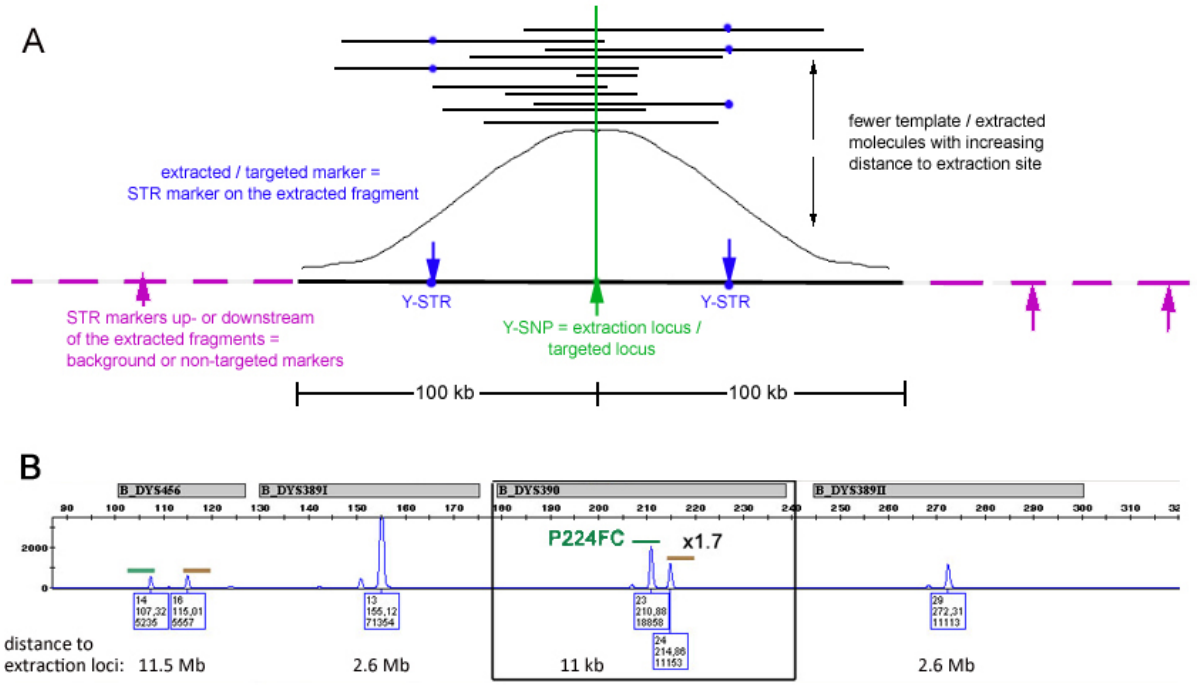


Figure 2-8: Overview of in this study used designation.
(A**) Y-chromsomal STR markers, which were aimed to be extracted, are designated as targeted side or extracted loci and are located on the extracted fragments (blue). Only SNPs serve as extraction sites and present the hybridization site of the probe (green). Markers which could be analyzed but show long distance to the extraction locus (>100kb) are designated as background (signals), non-separated or non-targeted markers (purple).**
(B) Blue channel of the STR profile generated with the AmpFLSTR® Yfiler multiplex kit from a HSE sample. In this example HSE was carried out with the probe P224FC as extraction loci, which is in 11 kb distance to the extracted or targeted marker DYS390. All other markers show long distance to P224 and belong to the background. Section labeled with black frame equals the cut out of Figure 3-3 B. Dark green bars label alleles of contributor 1 (5573) and brown bars label alleles of contributor 2 (1118) (see chapter 3.2).

In contrast to the HSE success (chapter 2.14.4) background signals were calculated as mean peak areas for a set of non-targeted loci to determine the degree of non-specific extraction occurring during HSE. These non-targeted loci include the marker DYS19, DYS385, DYS391, DYS392, DYS393, DYS438, DYS448, DYS456, DYS458 and GATAH4. Except of DYS438, all of these markers were never aimed for extraction in this study. HSE reaction with the probe Tat, which targeted and separated the marker DYS438 were not included in background calculations. Peak areas for single markers were calculated based on:

a) if alleles of the two contributors were different:

$$\text{mean peak area} = \frac{\text{peak area of allele one} + \text{peak area of allele two}}{2} \qquad (15)$$

b) if alleles of the two contributors were the same:

$$\text{mean peak area} = \frac{\text{peak area of allele one or two}}{2} \quad (16)$$

c) if one allele dropped out:

$$\text{mean peak area} = \frac{\text{peak area of allele one (+0)}}{2} \quad (17)$$

d) if both contributor alleles showed a dropout, the mean peak area was set = 0. Dropouts were defined based on a cut-off peak area of 200 and were included in the mean peak area calculation. The cut-off of 200 peak area refers to the laboratory evaluated standard background value of the Y-filer STR analysis.

e) For the calculation of the average value of the entire background of the obtained Y-filer electropherograms, the mean peak area values (calculated based on formulas 15-17) of the 10 Y-filer markers were averaged by:

$$\text{average background} = \frac{\begin{array}{c}\text{mean of DYS456, DYS458, DYS19, DYS385, DYS393,} \\ \text{DYS391, DYS392, DYS438, DYS448, GATA H4}\end{array}}{10} \quad (18)$$

Due to stochastic effects during PCR as well as small variations during capillary electrophoresis (e.g. different amounts of laser power during detection or differently aged capillary arrays), the peak areas exhibit fluctuations, even for identical samples. Therefore mean peak areas were presented in boxplot diagrams. For illustration, three representative markers of the Y-filer set (DYS19, DYS438, and DYS393) are shown. In this case, DYS19 is a marker with low average peak areas, DYS438 a marker with middle, and DYS393 with high peak areas.

2.15 Statistics

The arithmetic mean ($\bar{x}$) and standard variations (SD) were calculated with the programs Microsoft excel or PASW based on equation (19) and (20). The standard variation was calculated by the summation of the derivations of the different values x_i from the mean ($\bar{x}$). The sum was then divided by the number of experiments (n). In some cases SD was given also when the number of experiments was only two. In these cases, SD was calculated only to provide an orientation for the obtained variance. However, it should be considered that the calculation of SD from a small number of experiments appears rather conservative because variations are taken to the square, and therefore high variations are gaining more weight.

$$\bar{x}_{arithm} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad (19)$$

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n\text{-}1}} \qquad (20)$$

Significance was calculated according to the non-parametric Mann-Whitney Test with PASW, when one non-normal distributed variable has been compared for two independent samples. For example, HSE reactions which were carried out either with the standard protocol (sample A) or with two additional wash steps (sample B), should be compared by their obtained background signal intensity (peak areas) of the Y-filer STR profile. Using the Mann-Whitney test, obtained peak areas were set up as the non-normal distributed variable and HSE reactions (with or without two additional wash steps) were selected as the independent samples. The test compares all values of sample A with all values of sample B and calculates ranges as well as the asymptotic significance. The program uses a bidirectional test with the null hypothesis of A=B and an alternative hypothesis of A<B or A>B. The null hypothesis is rejected when p>0.05 with a confidential interval of 95%.

3 RESULTS

3.1 Search for variances close to forensic relevant Y-chromosomal STR-markers

HSE derives its power from the discrimination of single nucleotide polymorphisms (SNPs) which are placed at the 3'-terminus of the probe and create a mismatch position. Therefore SNPs or other short variances suitable for specific probe design were searched on distinct loci near Y-chromosomal forensic relevant STR-markers. Hereby, the selection was focused on STR markers, which are included in the Y-filer kit from Applied Biosystem, since this multiplex PCR presents one of the standard application in forensic for Y-STR detection (Figure 2-7). Furthermore, it was also searched for polynucleotide polymorphisms (e.g. di- or trinucleotide polymorphisms or short indels) in order to increase the number of mismatches at the 3'-end of the probe.

The starting point of the SNP search was the NCBI database because it covers most available SNPs, supplies additional information (e.g. submitter details, mapping position and allele frequencies) and contains convenient search tools. Further database searches were carried out with the International Hap Map Project, 1000 Genome Project, YCC, YHRD, and the Family Tree database. Most of the SNPs with ambiguous mapping could be excluded by comparing their positions in the different databases.

3.1.1 Database search

In order to identify polynucleotide polymorphisms, a SNP search was carried out in three different Y-chromosomal regions which include the Y-filer STR-marker DYS391, DYS635, DYS437, DYS439, DYS389I+II, DYS390 and span together 490 kb. In those regions of the Y chromosome the NCBI database showed 512 SNP entries, of which 58 entries had ambiguous mapping, 67 entries were described as "valid" and five entries were submitted as polynucleotide polymorphisms which were the primary focus. The rest of the entries showed neither ambiguities nor a valid status. Only three of the five polynucleotide polymorphisms (rs35753737, rs34485380, rs33963329) were chosen for further investigation. Here, BLAST analysis did not show any matches in the NCBI sequence database. This excluded the possibility that they might be paralogous sequence variations. Due to the absence of any homologous regions in these three variations, only the Y chromosome loci were tested.

In addition to the dbSNP NCBI database, the Insertion/Deletion database of the Mammalian Genotyping Service was used for the search for further polynucleotide polymorphisms. This search was carried out for all STR marker loci included in the Y-filer kit and provided one additional marker,

MID504, located close to DYS393, which was not submitted to the NCBI database. Because this entry included a comprehensive population study (Figure 2-1), the MID504 region was typed even though BLAST analysis of the MID504 locus showed a duplicated region on the X-chromosome. Therefore, an X-chromosomal specific primer was designed for the analysis of the paralogous X-chromosomal sequence of 19 male- and 20 European female samples. Furthermore, the SNPs rs2757249, rs2534929, rs2534930, rs2534931 and rs5940712 were sequenced as well because they are located in the neighborhood of MID504 and are included in the amplified fragment of MID504 (Figure 3-1). For all assays the primers were designed based on BLAST results to prevent potential additional PCR products originating from homologous regions. Y-chromosomal primer specificities were tested by an additional control PCR reaction with female DNA to confirm that in the absence of Y-chromosomal DNA no PCR product could be observed (Figure 2-2C, Figure 3-2).

The SNP entries rs35753737 and rs34485380 were investigated by the sequencing of 65 and 68 Y chromosomes, respectively. None of the tested Y chromosomes showed any variance, and only the allele that corresponds to the reference sequence was detected. The same results were obtained for the locus rs33963329, where 64 chromosomes from a laboratory DNA library were tested. Except for rs5940712, none of the analyzed database entries in the MID504 region showed a variation on the X or Y chromosomes. The sequence difference between the chromosomes at the MID504 locus (X chromosome: ACA present, Y chromosome: ACA deleted) was identical to the base substitutions of the database entry (ACA/-). The same was observed for the eight additional entries on the alignment fragment (Figure 3-1). A summary of all results obtained either by Pyro- or by Sanger sequencing are shown in Table 3-1 and (Appendix-Table 7-1, Appendix-Table 7-2).

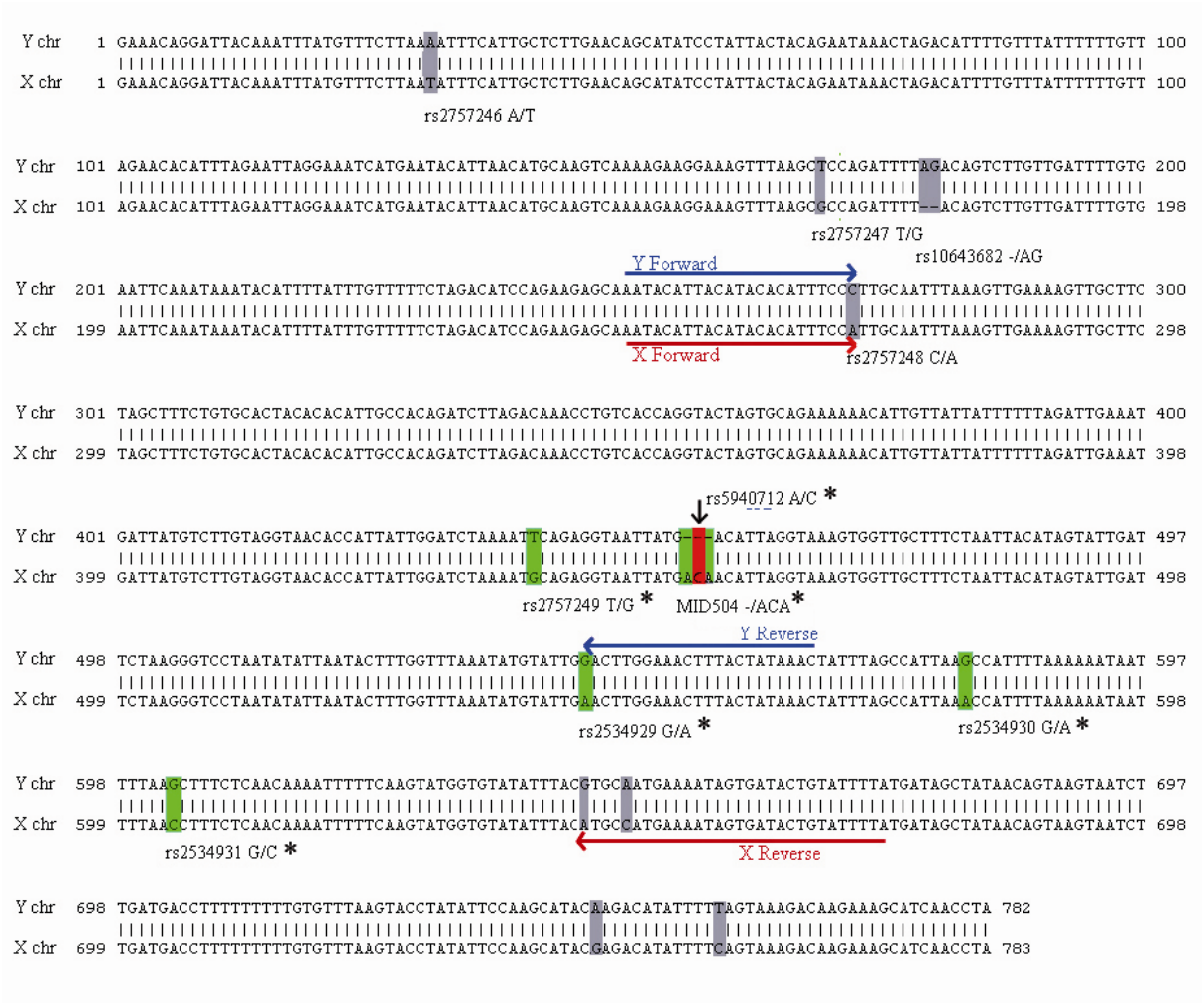


Figure 3-1: BLAST result of a 782/783 base sequence region encompassing locus MID504.
The MID504 and adjacent regions on chromosome Y were aligned with a corresponding sequence on the X chromosome in the human sequence database. All differences in sequences between X and Y fragments, including MID504, are marked with shaded boxes and, when available, the rs number and base substitution from the NCBI database entries are given below. Asterisks indicate SNPs analyzed in this study (green boxes). The vertical arrow highlights the single base variance of rs5940712 (red box). The horizontal arrows show the positions of Y (blue) and X (red) specific primers used for sequencing analyses in this study.

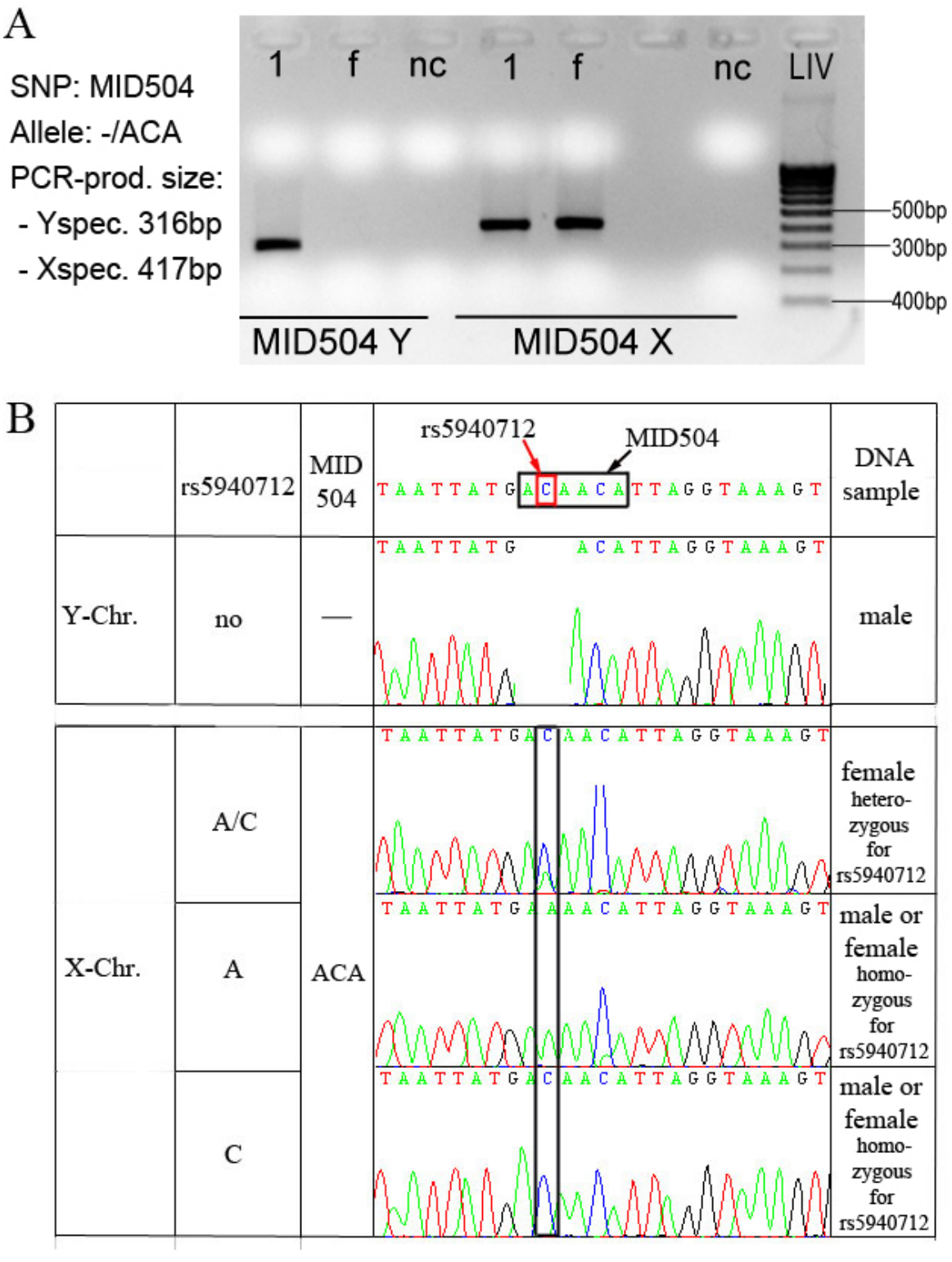


Figure 3-2: Typing of MID504. (A) Y-chromosomal specificity of PCR before Sanger sequencing.
PCR reactions were controlled on a 1.4% agarose gel at 100 V for 30 min. Analysis of MID504 was carried out with Y- and X-chromosome specific PCR-primers. Primer-specificity for the Y-chromosome was tested with male- (1) and female-DNA (f). (nc) negative control, (LIV) HyperLader IV. (B) Overview of observed sequence results for the MID504 locus on the Y- and X- chromosome as well as for male and female DNA. Note that inside the ACA insertion on the X- chromosome, a second high polymorphic SNP, rs5940712, occurs.

Single nucleotide polymorphisms were mainly selected from the Y-chromosomal phylogenetic tree, published by Karafet *et al*. 2008 [142]. The phylogenetic tree contains over 600 polymorphisms of which the majority presents highly valid, biallelic SNPs with known heterozygosity for different populations. After the screening of a dataset containing 577 unique mapping phylogenetic variances, 85 SNPs were found to be located close to Y-filer STR markers of which 35 were situated between the marker group DYS635 and DYS389. Taking into account the typical heterozygosity for European populations, nine phylogenic markers (P29, P30, P38, P40, P41, P224, P240, P244, and Tat) were selected as a potential HSE extraction locus. In addition also the SNPs M343, M198, and M173 were selected as extraction loci, which showed larger distances (243, 93, and 88 kb) to the next respective Y-filer STR marker as well as the two validated, non phylogenetic SNPs from NCBI database, rs13304202 and rs2040607. The last two belong to the 67 valid entries found in the search for polynucleotide variances and are also located in the marker group DYS635 to DYS389. As before described, SNP typing was carried out with Y-chromosomal specific PCR prior to Pyro- or Sanger sequencing.

All selected SNPs were typed with a relatively small male DNA sample set (up to 26 samples) but showed high heterozygosity, with the exception of P29 and P41.2. As expected, the derived allele for the SNP Tat was found only in one sample with north-European ancestry. No variation was found for the SNPs P29 and P41.2, concluding that the selected sample set did not comprise haplogroups E and I2a1. Therefore, P29 and P41.2 were not available for further HSE-probe design. The SNP rs2040607 from the NCBI database showed a different genotype only once, but belongs to a sample of which only insufficient DNA material for further HSE applications was available. Therefore, rs2040607 could not be used as an HSE-extraction locus but the here obtained typing result confirmed the NCBI entry. A summary of all obtained results are listed in Table 3-1.

Table 3-1: Summary of all SNPs analyzed in this study and selected database information.

name[a]	position[b]	allele-frequence[c]	validation status[d]		allele[f]		result on Y Chr.		result on X Chr.		BLAST results indicate PSV	possible reason for false SNP entry
			methode	sample size	A 1 (anc)	A 2 (der)	A 1	A 2	A 1	A 2		
P29 - E	13007215 ARSDP	Ymap 0,074	Pop	408	A	C	**19**	**0**	-	-	no	n.v.i.s.
P30 - I1a	13006761 ARSDP	Ymap 0,446	Pop	836	G	A	**15**	**6**	-	-	no	true
P38 - I1	12994387 ARSDP	Ymap 0,188	Pop	473	A	C	**15**	**6**	-	-	yes	true
P40 - Ia1	12994402 ARSDP	Ymap 0,09	Pop	455	C	T	**15**	**6**	-	-	yes	true
P41.2 - I2a2a	13001679 ARSDP	Ymap 0,004	Pop	554	T	C	**20**	**0**	-	-	yes but only 87% seq.cov.	n.v.i.s.
Tat - N1c (rs34442126)	13431977 USP9Y	MIX 0,021	Pop	169	T	C	**5**	**1**	-	-	no	true
M170 - I (rs2032597)	13357186 USP9Y	EU 0,692	Pop	1138	A	C	**6**	**1**			no	true
M173 - R1 (rs2032624)	13535818 DDX3Y	EU 0,538	Pop	797	A	C	**7**	**6**	-	-	no	true
M198 - R1b1b2g1 (rs2020857)	13540146 NLGN4Y	EU 0,077	Pop	2405	G	A	**17**	**3**	-	-	no	true
P224 - R1 (rs17307398)	15795387	EU 0,538	Pop	723	C	T	**5**	**4**	-	-	yes but only 86% seq.cov.	true
P240 - P (rs6530605)	13108816 XG	EU 0,538	Pop	176	T	C	**11**	**14**	-	-	no	true
P244 - P (rs2740981)	12943108	EU 0.583	Pop	36	G	A	**11**	**14**	-	-	no	true
M343 - R (rs9786184)	2947824	Ymap 0,607	Pop	713	C	A	**12**	**9**	-	-	no	true
rs13304202	12926224	EU 0.667	Affymetrix	525	G	A	**11**	**10**	-	-	no	true
rs2040607	13039638	MIX 0,74	Sanger Sequencing	184	A	G	**25**	**1**	-	-	yes	true
rs35753737	12579277 - 12579278	-	Celera	Celera Donors[e]	TA	AT	**65**	**0**	-	-	no	MAF
rs34485380	13104633 - 13104634	-	Celera	Celera Donors[e]	TC	GT	**68**	**0**	-	-	no	MAF
rs33963329	15774238 - 15774239	-	ReSeq	2	DEL	TTA	**64**	**0**	-	-	no	repeat
MID504	3190358 - 3190360	EU 0,72	Pool	430	DEL	ACA	**48**	**0**	0	59	yes	PSV
rs2534929	3190368	-	BAC	2	G	A	**18**	**0**	0	44	yes	PSV
rs2534930	3190405	-	BAC	2	G	A	-	-	0	31	yes	PSV
rs2534931	3190428	-	BAC	2	G	C	-	-	0	51	yes	PSV
rs2757249 v	X Chr. 88747775 suspected	-	BAC	2	T	G	**18**	**0**	0	46	yes	PSV
rs5940712 v	X Chr. 88747791	EU 0,126	BAC	396	C	A	-	-	44	15	no	true SNP on X

The white side (left) shows data taken from public database and the orange side (right) shows results from this study.

a) SNPs with rs numbers refer to dbSNP database entries, SNPs designated with P, M, and L numbers are taken form Family Tree DNA database, and MID number is from the Mammalian Genotyping Service. When a SNP belongs to the phylogenetic tree, its haplogroup is given behind the name.

b) The Y-chromosome position is given after the reference sequence hg 18. If a SNP is located in a putative gene, the name is given below. The SNPs rs2757249 and rs5940712 are located on the X chromosome (X Chr.). Since June 2011 rs2757249 is signed as suspect in the NCBI database but showed validation status (v) at the time of this study.

c) Allele-frequencies are taken from the NCBI database (EU = European Population; MIX = different ethnic populations) or are derived from given counts in the Family Tree database (Ymap).

d) Type of SNP analysis method and sample size are used to describe the validation status: Pop = SNP belongs to Y-chromosomal phylogenetic tree and tested in population studies [142], Affymetrix = genotypes were produced by the International HapMap project Consortium, by the AFFYMETRIX genotyping center using the affymetrix-genome-wide-snp-6.0 platform, Ymap = SNP was tested by Family Tree Database, Celera = whole genome shotgun assembly (WGSA) of short fragments and public bacterial artificial chromosome contiguous sequences (BACtigs) assembled with Celera software [168, 169], ReSeq = assembly of re-sequenced traces from the SNP discovery project [170], Pool = pooled PCR discovery [171], BAC = BAC overlap discovery [172].

e) Celera Donors included two males and three females - one African-American, one Asian-Chinese, one Hispanic-Mexican and two Caucasians.

f) Point mutations are shown as allele 1 (A1) and allele 2 (A2) for forward strand.
For all phylogenetic SNPs, allele 1 presents ancestral (anc) and allele 2 the derived (der) status. (-) no data, (MAF) minor allele frequency, (PSV) paralogous sequence variant, (DEL) deletion, (n.v.i.s.) SNP showed no variation in this study.

3.1.2 Search by sequence analysis

With the perspective to discover a polynucleotide variance and to obtain potential HSE extraction sites very close to one STR marker 4.8 kb of sequence, which flanking the DYS437 marker were sequenced for ten DNA samples. For this purpose, six overlapping primer-pairs (S1-S6) were designed to amplify the entire sequence (Table 2-3, Appendix-Figure 7-2). Although the selected 4.8 kb sequence did not show any BLAST result outside the Y-chromosome, some unspecific products were observed for fragment S2 and S3 for both male and female samples. However, sequencing results did not appear to be influenced by these unspecific products because subsequent sequencing reactions were carried out with nested sequencing primers and for female samples no or inconclusive sequences were obtained (data not shown). The sequencing results revealed two new single nucleotide variations for two different samples: a C to T transition for sample 110776 at position 2701 (S4:2701) and a G to C transition for sample 94447 at position 4204 (S6:4204). Furthermore, the two non-validated NCBI database entries (rs35938675 and rs35938675), which also map in the 4.8 kb sequenced fragment, were not found as a polymorphic in the ten investigated samples.

3.2 First separation tests of male DNA mixtures by HSE

STR analysis of DNA samples that contain more than one contributor result in mixed profiles. To test whether these mixed profiles can be resolved using HSE, first separation tests were carried out according to the HSE manual provided by Qiagen for the resolution of HLA loci. Subsequent to the standard extraction of male DNA mixtures, the obtained HSE samples were analyzed by PCR for the detection of extracted and non-extracted Y-STR markers.

3.2.1 First HSE results with the Y-chromosomal probes M173, M198, M343 and P224

For the first separation tests of male DNA mixtures, the Y-chromosomal SNPs M343, M173, M198 and P224 were selected as extraction loci because of their close distance to the Y-chromosomal STR markers DYS495 (20 kb to M198 and 15 kb to M173), DYS504 (15 kb to M343), DYS536 (25 kb to M343), and DYS390 (11 kb to P224). DNA mixtures were created by mixing equal amounts of DNA from two known male contributors who possess different alleles at the selected markers. For example, contributor 1 (5573) shows the allele G at the extraction site P224 and the allele 23 for the marker DYS390, whereas contributor 2 (1118) carries the alleles A and 24 (Figure 3-3A). To separate Y-chromosomal segments of the generated DNA mixture (contributor 1 /contributor 2), HSE was tested with probes that were specific for the respective extraction loci. Furthermore, HSE was done without probes as a control extraction. HSE was conducted as described in chapter 2.14 with 1x HB which was included in the HaploPrep Kit from Qiagen and one of the following probes: P224 FC/FT, RG/RA (23 nt long), M173 RT/RG (24 nt long), M198 RA/RG (25 nt long) and M343 FC/FA (20 nt long). Finally, HSE samples were analyzed by the detection of the extracted Y-STR markers by manually set up singleplex PCRs or for the marker DYS390 with the AmpFLSTR® Yfiler. HSE success was then evaluated by calculating the ratio between the contributor alleles or the enrichment of one contributor as percentage. Results are illustrated as electropherograms in Figure 3-3B. For example, the electropherogram of the control extraction (no probe) of locus DYS390 shows a mixed profile with an equal portion of the allele peak areas 23 (contributor 1) and 24 (contributor 2), which reflects the 1:1 DNA mixture generated at the beginning.

When HSE was conducted with probes specific to the marker P224, the corresponding allele of the marker DYS390 appeared to be enriched to an average of 1.5-fold in the detected profile. Hereby, an enrichment of the allele 23 of contributor 1 could be observed when using the corresponding probes P224 FC (forward) and P224 RG (reverse). On the other hand, when using the probes specific to allele A, which correspond to contributor 2 (P224 FT and P224 RA), the allele 24 of the marker DYS390 could be extracted. An enrichment of one contributor could also be detected for HSE reactions with the probes M173RA/RC and M198RT/RC (61% till 64%), whereas HSE with the probe M343 showed no separation (data not shown). The detection of the marker DYS495, DYS504 and DYS536 showed low sensitivity (low peak highs) or enhanced stutter signals, which complicated HSE analysis. Therefore, further analysis of HSE samples were done only with the AmpFLSTR® Yfiler multiplex kit.

The extraction loci M343, M178, and M198 are also located near one STR loci of the Y-filer kit but have larger distances of 88 kb to 243 kb. Analysis of the Y-filer STR marker after the extraction

reaction with M343, M178, and M198 did not show any indication of enrichment (data not shown). In summary, the first HSE experiments showed low or no separation effect for all tested probes.

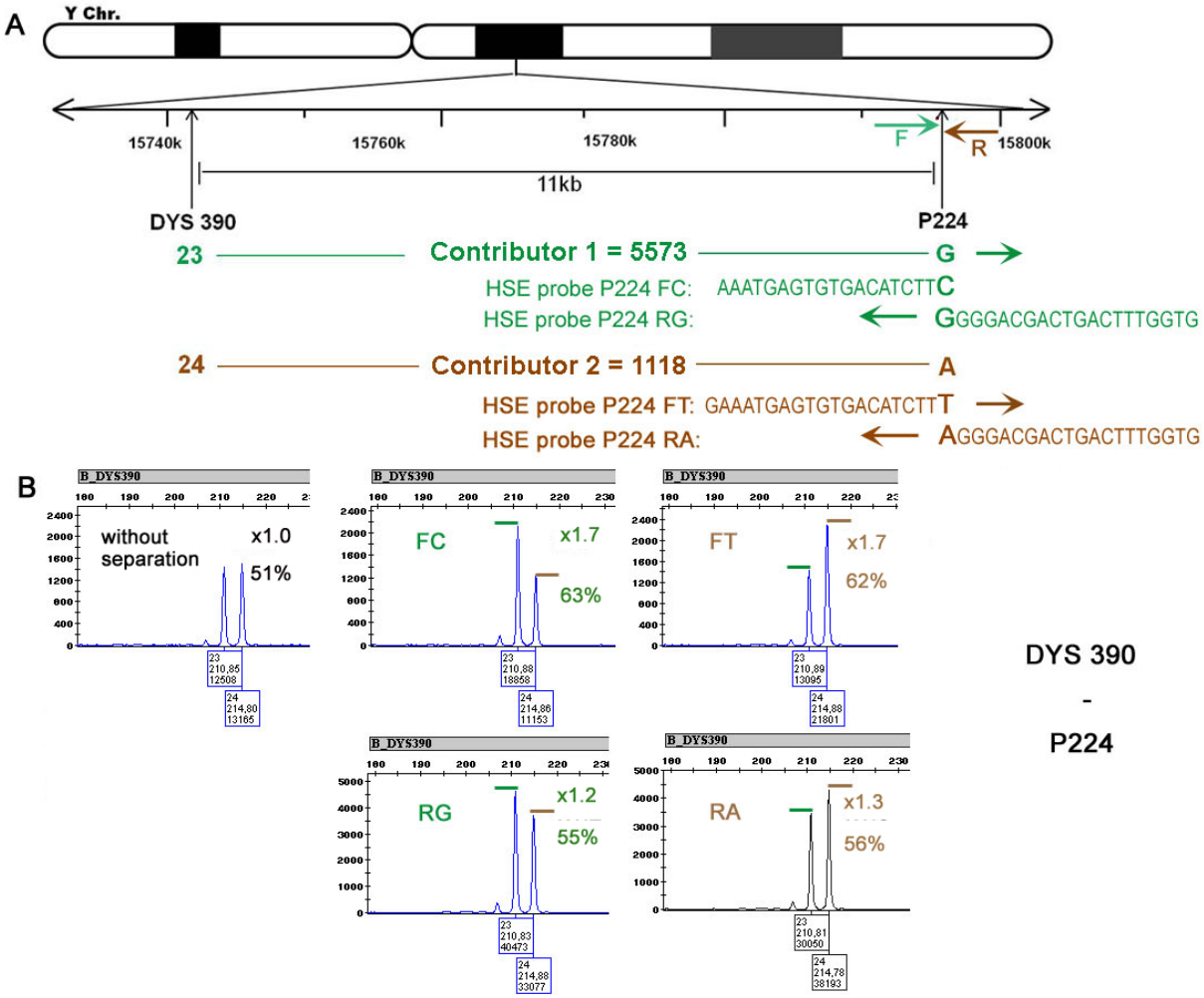


Figure 3-3: First separation tests of male DNA mixtures by HSE at different Y-chromosomal STR loci.

(A) Schematic map of the design of HSE reactions at the example of the two loci DYS390 and P224, which are separated by a distance of 11 kb and located on the human Y-chromosome. Arrows indicate the positions of the specific probes, which are labeled with F (forward), R (reverse), and the 3'-termini-specific nucleotide. The chromosome positions given are based on the NCBI reference sequence (Hg18).

(B) Electropherograms show STR alleles obtained from HSE reactions without probe or with probes specific to one of the contributors of the DNA mixture. The locus DYS390 was detected with the AmpFLSTR® Yfiler kit. Separation success of the DNA-mixture by HSE is given as ratio and enrichment of one contributor as percentage for each probe. The green color corresponds to alleles, probes, and enrichment of contributor 1 (sample ID 5573) and brown color to those of contributor 2 (1118). Sample ID numbers refer to Table 2-8.

Surprisingly, Y-STR analysis of DYS390 with the Y-filer kit showed also strong signals for all alleles of both contributors in all other non-separated Y-filer markers, which appeared to equal portions and reflected the 1:1 DNA mixture. In this study non-separated Y-STR markers refer to markers, which are located outside of the extracted Y-chromosomal fragments, whereas targeted-loci are located within the extracted fragment (Figure 2-8).

3.2.2 Investigation of the HSE background

First HSE tests with male DNA mixtures did not show high enrichments of extracted Y-STR markers but showed strong background signals in Y-filer multiplex PCR analysis. In addition the alleles of the two contributors of the DNA mixture could be detected also in HSE without any extraction probe (chapter 3.2.1). Further investigations of HSE samples with autosomal STR typing kits (SE-filer and Identifiler of Applied Biosystems) also resulted in complete profiles (data not shown). The appearance of these background signals were unexpected and could potentially arise from unspecific binding of DNA to the streptavidin-beads, which also could in turn worsen the separation effect. Therefore, the Y-filer PCR background detected after HSE reaction was investigated under following aspects:

- Different amount of DNA in HSE reaction
- Additional washing of the streptavidin-beads
- Additional blocking of the streptavidin-beads
- Influence of different HSE buffer components

All in this chapter 3.2.2 described HSE reaction were carried out with the probes P224 FC/FT, RG/RA (23 nt long) or the probes M173 RT/RG (24 nt long).

3.2.2.1 Influence of DNA amount on HSE background signals

In this study, standard HSE reactions were carried out with approximately 300 ng of DNA mixture. After an extraction reaction, the HSE sample was eluted in 50 µl elution buffer of which 7.5 µl contained sufficient DNA for direct Y-filer analysis.

In order to understand if the observed background arises through an overload of the HSE reaction with DNA, HSE was tested with different amounts of input DNA. Therefore starting DNA mixtures were created by mixing equal amounts of male DNA and then measured with the spectrophotometer. Next, mixtures were serially diluted and each dilution was measured with the RotorGene Q instrument. HSE was carried out in duplicates for each amount of DNA input (50 ng, 20 ng, 10 ng, 5 ng, and 1 ng) according to the standard protocol (chapter 2.14). Results show that, as expected, the detected background signals from HSE samples could be decreased with less input of DNA (data are presented together with data from chapter 3.5.3 in Figure 3-17). Hereby, almost no signals were obtained for 1 ng input DNA and signals about 1000 peak area for 5ng input DNA.

However, no improved enrichment of the extracted locus could be detected for HSE reactions with decreased amounts of DNA mixtures.

3.2.2.2 Influence of additional blocking and washing steps on HSE background signals

Additional washing- and blocking steps were included in the HSE protocol to test whether high HSE background signals occur through unspecific binding of DNA to the streptavidin-beads. The streptavidin beads were removed from the Qiagen-cartridges and incubated with high amounts of human female DNA (Figure 3-4A). The beads were tested for successful blocking by washing them once (Figure 3-4B) and then detecting the female STR profile by the Identifiler Kit. The electropherograms obtained from the resuspended bead-solution (see Figure 3-4C-E) showed complete profiles of the female DNA with an average peak area of 1300. Therefore, it was concluded that unspecific binding of DNA to the beads indeed had occurred.

Furthermore, additional washing steps were tested, too. For this purpose, beads which were incubated with female DNA as described above were washed manually two more times (Figure 3-4F) and used again for autosomal STR detection by the Identifiler Kit (Figure 3-4G-I). However, the STR analysis of these beads showed a strong reduction of the desired signals in the electropherograms, with an average peak area under 200, which is under the cut off value for background signals (see chapter 2.14.4) and a dropout rate between 100% and 65%.

Therefore, it can be conclude that first DNA can bind unspecific to the beads, and second this unspecific bound DNA can be removed by increased washing. In order to test if the observed HSE background can be decreased by a more intensive wash, the HSE standard protocol was carried out with two additional washing steps and analyzed with the Y-filer multiplex PCR. However, the results showed no significant reduction of the HSE background and no improvement of the HSE separation effect (data not shown).

A reduction of the background signals or and improved separation was also not found for HSE reaction, which were carried out after standard protocol but with blocked beads (data not shown).

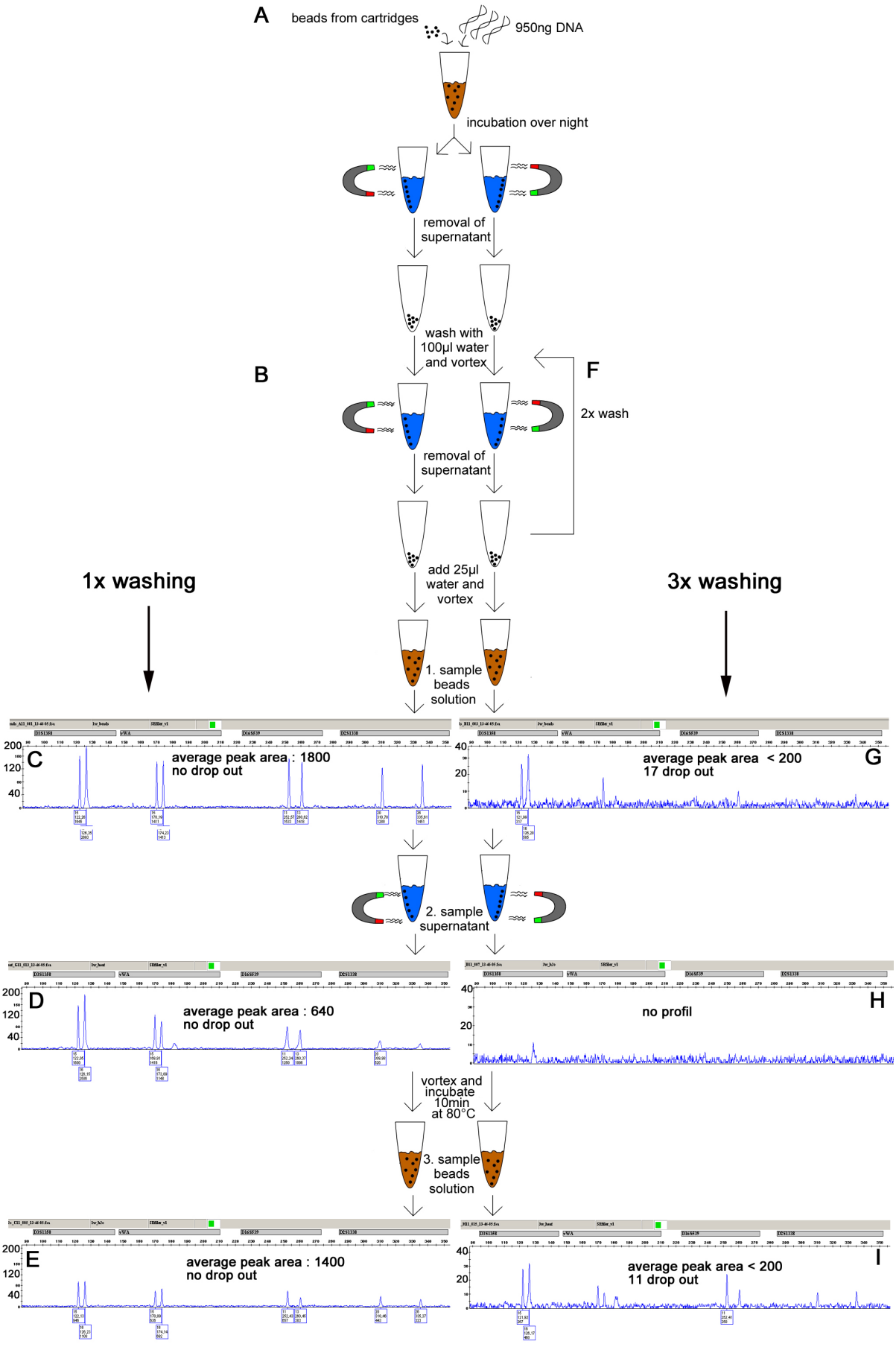

A
beads from cartridges
950ng DNA
incubation over night
removal of supernatant
wash with 100µl water and vortex
B
F
2x wash
removal of supernatant
add 25µl water and vortex
1x washing
3x washing
1. sample beads solution
C
average peak area : 1800
no drop out
G
average peak area < 200
17 drop out
2. sample supernatant
D
average peak area : 640
no drop out
H
no profil
vortex and incubate 10min at 80°C
3. sample beads solution
E
average peak area : 1400
no drop out
I
average peak area < 200
11 drop out

Figure 3-4: Demonstration of unspecific binding of DNA on streptavidin beads and subsequent removal by washing. (A) Female DNA is added to beads solutions and incubated over night. (B) After incubation with female DNA, beads are fixed in the tube by a magnet and supernatant, containing female DNA, is removed. Beads are then washed once and wash solution again is removed. After wash, beads are resuspended again in water for analysis with the Identifiler multiplex PCR by using either (C) resuspended beads-solution containing beads, (D) resuspended solution without beads or (E) resuspended beads solution containing beads after an additional heating step of 80°C for 10 min. (F) Washing step from (B) was repeated two times more before analysis. Results of additionally washed beads are shown in (G) to (I) and were obtained after the same workflow of (C) to (E).

3.2.2.3 Influence of single components of the hybridization buffer on the HSE background

The HSE HB consists of four basic components: PCR buffer, dNTPs, biotin-dUTPs and the polymerase. In order to investigate the influence of the single buffer components on the Y-filer HSE background noise, HSE was carried out with the complete removal of one or more HB components. Therefore different HB were set up and HSE reactions were then carried out with the modified HBs according to the standard protocol (chapter 2.14) and 300ng input DNA. Next, HSE samples were analyzed with the Y-filer kit and compared based on their obtained profiles. Results show that low or no background signals could be detected for all reactions without polymerase or biotin-dUTPs. HSE with complete HB or those missing only dNTPs however exhibits very strong and complete profiles at all loci.

Table 3-2: HSE background signals from HSE with different buffer compositions.

type of buffer	content of used buffer	background	n
complete buffer	1x PCR buffer, polymerase, dNTPs and biotin-dUTPs	**+**	8
no dNTPs	1x PCR buffer, polymerase, and biotin-dUTPs	**+**	2
no biotin	1x PCR buffer, polymerase, dNTPs and probe	**-**	5
no polymerase	1x PCR buffer, biotin-dUTPs, dNTPs and probe	**-**	2
no HB	only water or 1x PCR buffer	**-**	7

(+) Y-filer analsis showed strong HSE background, (-) Y-filer analysis showed low or no HSE background, (n) number of experiments.

The influence of the polymerase and the biotin on the HSE background could be also seen when HSE reactions were carried out with different concentrations of polymerase and biotin-dUTP (see chapter 3.3.2). Here HSE with decreasing polymerase or biotin-dUTP concentrations indicated a decrease of the background signals ($p_{polymerase}$= 0.007, p_{biotin}= 0.014, data not shown). On the other hand, an increase of these components did not show higher background.

3.3 Optimization of the HSE hybridization buffer

First tests of the separation of Y-chromosomal STR loci of male DNA mixtures by HSE resulted in only low enrichment of one allele and unexpected background signals. Previous results indicate an influence of HB components on background signals. Therefore the HB was to be investigated in more detail to achieve a more definite separation of one contributor's allele from DNA mixtures.

3.3.1 Test of different HB dilutions

First, different dilutions (½, ¼, and ⅛) of the HB of Qiagen were tested to determine whether any changes in the HB concentration can improve HSE separation. Results were evaluated by the calculation of the ratios of the peak areas (alleles 23 and 24) of the two contributors (Figure 3-5A). A ratio of one would reflect the same presence of two allele peaks, whereas a ratio greater than one reflects the enrichment of one contributor's DNA. HSE results using the diluted HB showed that the separation effect increased when HB is diluted up to 4-fold. No separation effect was observed when the HB was diluted 8-fold, presumably because the HB was too diluted (Figure 3-5A).

3.3.2 Test of different concentrations of single HB components

Due to the improved separation effect with diluted HB, further investigations were conducted with increasing and decreasing concentrations of the key HB components: the biotin-dUTPs, dNTPs, and *Taq* polymerase. For the different tested biotin-dUTP concentrations (Figure 3-5B) as well as for the different dNTP concentrations (Figure 3-5C), a significant improvement of the separation effect could not be detected. Biotin-dUTP showed a broad range of concentrations in which a specific extraction could occur (from 3x to 0.375x). With an increased concentration of 5x the separation effect was clearly lost (Mann Whitney Test; $p = 0.001$). The different dNTP concentrations tested resulted in a gradual loss of the separation effect. Furthermore, as observed for biotin-dUTP, the evaluation of different polymerase concentrations showed that the calculated ratios were decreasing with higher *Taq* polymerase concentration. However, at lower *Taq* polymerase concentrations the calculated ratios and therefore the separation effect increased (Figure 3-5D). The average of the measured ratio rose from a 1.5-fold increase for 1x *Taq* polymerase concentration to a 3-fold increase for ½x and ¼x concentrations of *Taq* polymerase (Mann Whitney Test; $p < 0.001$). HSE reaction with only ⅛x Taq polymerase concentration showed even a mean enrichment of 4-fold but with a broader variance. Here the use of minimal polymerase concentration could facilitate the occurrence of stochastic effects during HSE.

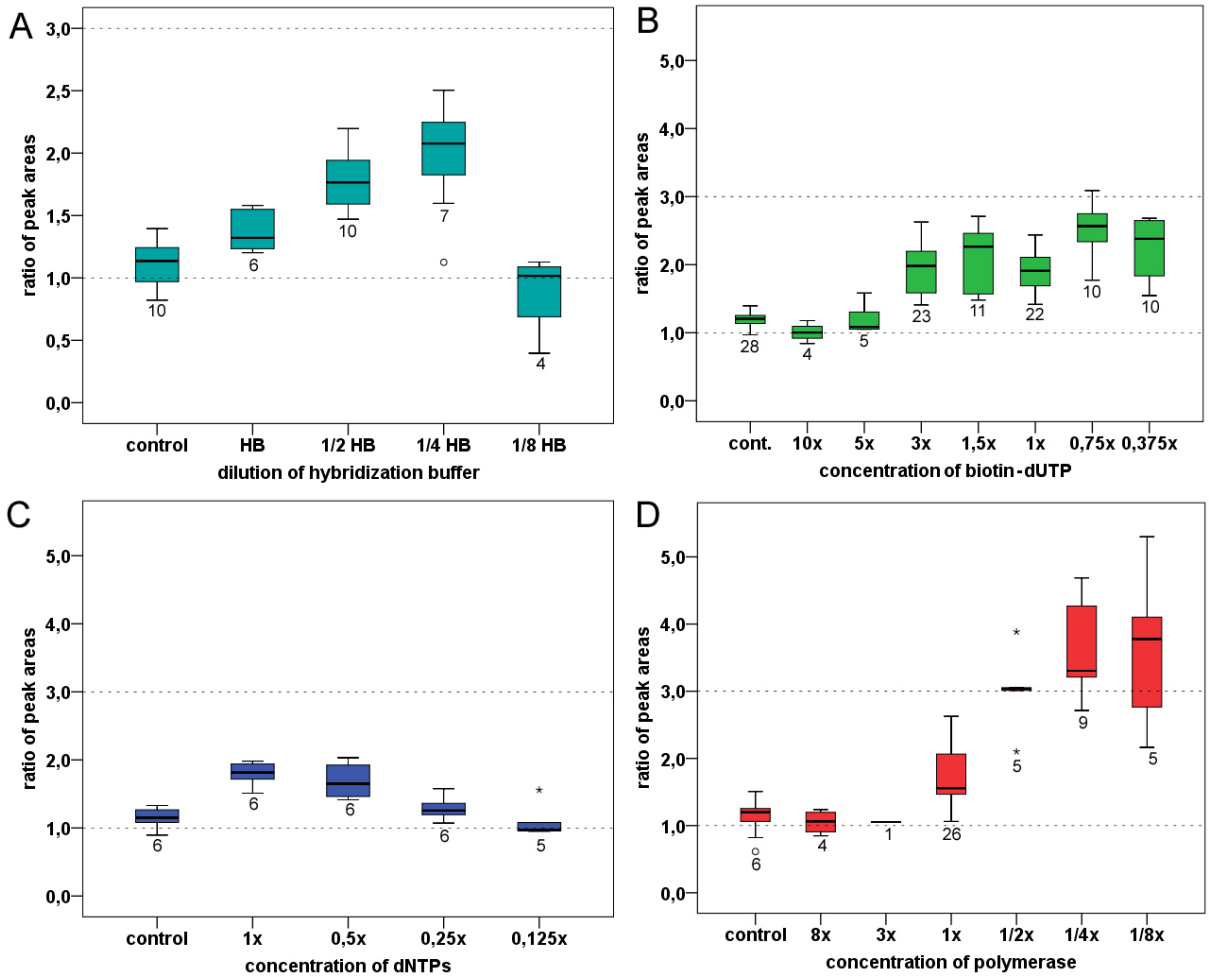


Figure 3-5: Separation effect obtained from HSE samples with different HB compositions.
Box blot diagrams (A-D) show the separation effect of the two alleles 23 and 24 of the locus DYS390 for different compositions of HB. Separation effect was calculated as ratios of the two contributor alleles. Given concentrations refer to the 1x hybridization buffer from the HaploPrepKit. For the HSE reaction, the probes FC, RG, FT, and RA were used randomly. Numbers represent the number of HSE reactions tested.

3.4 Optimization of HSE-probe specificity

The HSE separation effect could be improved through the optimization of the hybridization buffer but did not result in a complete separation or a pronounced appearance of the targeted allele. Therefore, further efforts were put into the testing of HSE-probe specificity, since the allele-specific extension of the hybridized probe by the polymerase is the basic concept of HSE.

3.4.1 Influence of probe length on HSE specificity

In previously described HSE optimization studies (chapter 3.2 and 3.3), only one probe was designed and tested (for the extraction loci M343, M173, M198 and P224). For example, for HSE at the

targeted locus P224, the probes P224 FC/FT and P224 RG/RA were tested for only one length (23 nucleotides, Figure 3-3A, Figure 3-6A).

In general, the design of HSE probes is limited by the specified targeted locus, which restricts design options to orientation and length. Therefore, new probe designs for P224 probes were tested with different length for each orientation. As described already before in chapter 3.2.1, HSE was carried out with one probe and a male DNA mixture containing two contributors that differ at both loci, DYS390 and P224. Next, Y-filer multiplex PCR was used to detect the DYS390 locus of the HSE P224 samples, which reflects the separation effect of the DNA mixture.

Electropherograms of the HSE samples show that HSE control reaction without a probe results in no enrichment of one contributor allele, whereas the use of probes increases the presence of one contributor allele. The 23 nucleotide-long probes led to an enrichment of 63%, 62% of the corresponding contributor as already observed in former studies (Figure 3-6). Surprisingly, the probe that was only four nucleotides shorter exhibited a strong increase in the separation of one contributor to nearly 100%. HSE with the probe P224 FT, 23 nucleotides long, result in an average enrichment of 60%, whereas the HSE success increased in 29% of enrichment, when probes were shortened by only three nucleotides. The orientation of the probe did not have any significant effect. A more comprehensive study of probe P224FC comprising tests of decreasing probe length (9, 21 to 23 nucleotides) revealed optimum HSE success with a length between 15 and 20 nucleotides, whereas longer or shorter probes led to a rapid decrease in HSE success. Similar results were also detected for the probe set P224FT (Figure 3-7A).

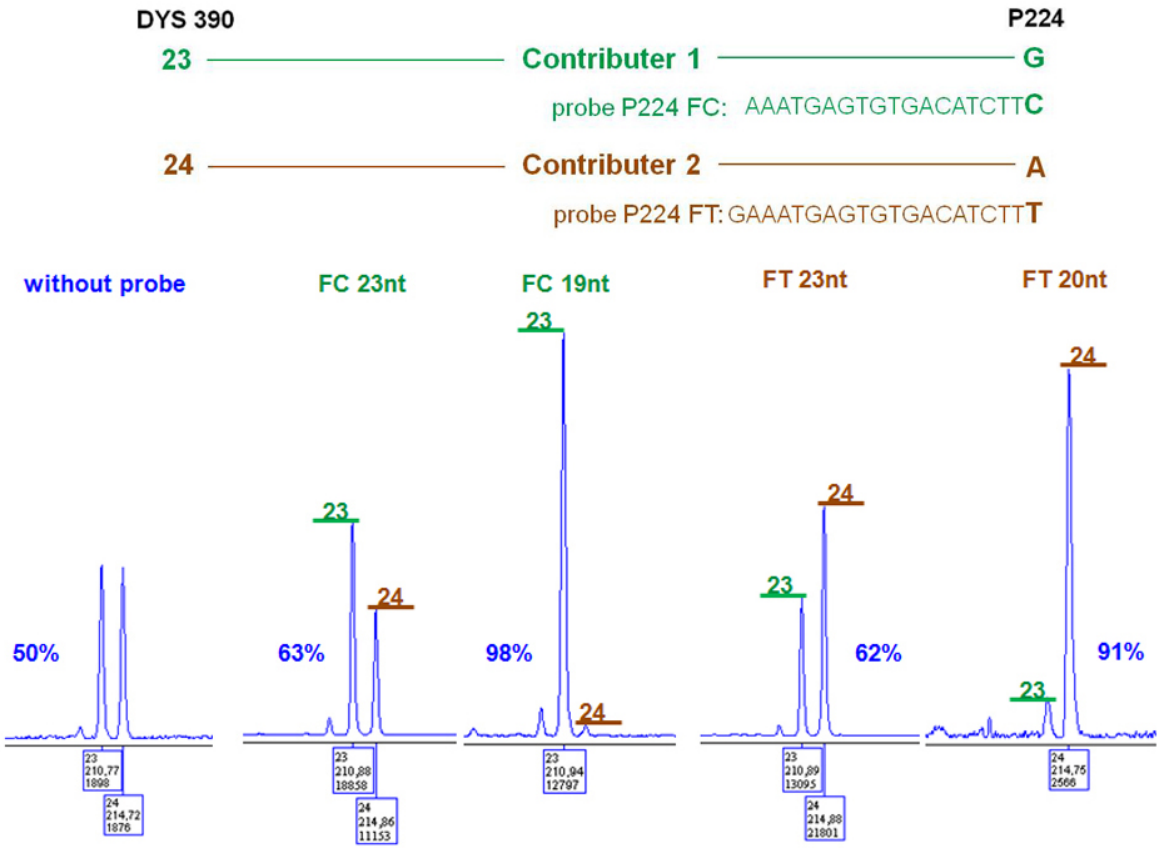


Figure 3-6: HSE separation with probes of different lengths.
Shown are the electropherograms from AmpFLSTR® Yfiler analysis of HSE samples, which were carried out with a DNA mixture (contributor 1 = 5573 and contributor 2 = 1118) and the probes P224FC (23 and 19 nucleotides long), P224FT (23 and 20 nucleotides long) or without probe. Green bars with allele 23 correspond to contributor 1 and brown bars with allele 24 correspond to contributor 2. Percentages indicate the enrichment of one contributor.

The newly designed probes for P224FC were also tested in HSE with the optimized hybridization buffer, which contains a reduced polymerase concentration (chapter 3.3.2). In this case too, the results showed an optimum curve of the HSE separation effect, with the best enrichment for probes P224 FC between 18 and 21 nucleotides long. A comparison between HSE reactions with standard and optimized buffer showed that first, the separation effect for extractions with probe 23 nucleotides long increased from 68% up to 82% enrichment (see chapter 3.3.2). Second, significant HSE separation occurs within a smaller window of probe length for HSE with optimized buffer. Here, the obtained enrichment of one contributor starts to drop beginning with a probe length of 17 nucleotides, as compared to a probe length of 14 nucleotides when carrying out HSE with standard HB. The probes P224FC, which are 17 and 16 nucleotides long, still achieved a complete separation of approximately 90% enrichment when are used in HSE with standard puffer, whereas the same probes result in an enrichment of only 77% when used in HSE with decreased polymerase concentration. The comparison of HSE with the probes 14 and 13 nucleotides long showed still a significant enrichment of 77% for HSE with standard puffer, but no significant enrichment for HSE

with decreased polymerase concentration. In summary it can be conclude that HSE success can be improved either with decreasing polymerase concentration or with shorter probe length. Furthermore, HSE with less polymerase concentration achieve an optimized separation for longer HSE probes but failed to maintain HSE success for shorter becoming probes.

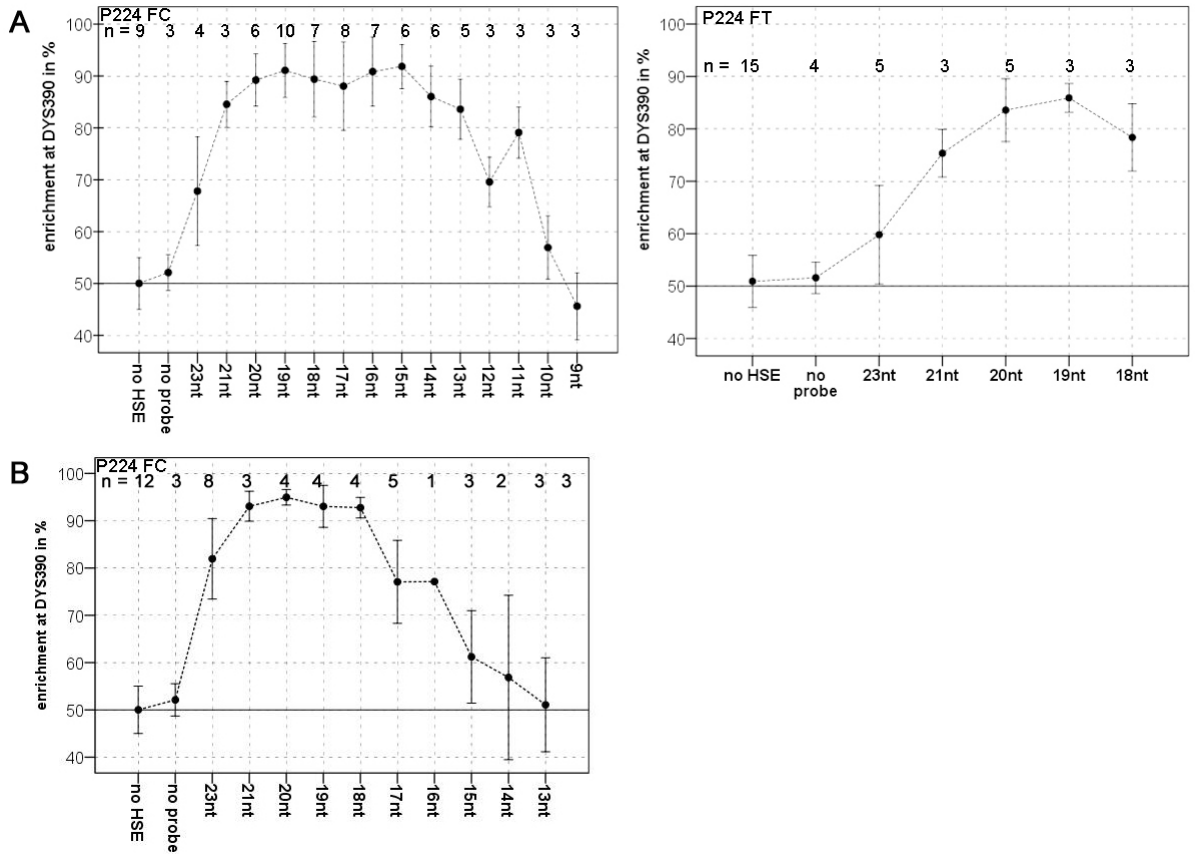


Figure 3-7: HSE success with different probe length of P224FC/FT.
(A) Enrichment of one contributor of a male DNA mixture is shown in relation to probe length for the two different probe sets P224FC and FT. (B) HSE with probe set P224FC and reduced polymerase concentration. HSE was carried out with male DNA mix 1118/5573. The bars indicate standard deviation of the mean enrichment. "No HSE" indicates analysis of the male DNA mixture without separation by HSE and "no probe" indicates separation of a male DNA mixture by HSE without probe. n = number of extractions.

3.4.2 Development and validation of a new prediction model for HSE probe specificity

In previous experiments (chapter 3.4.1), a clear dependency of the probe-length and HSE success for the probes P224FC and FT has been observed. In order to confirm these results and to better understand the dependence of HSE on probe length, HSE analysis were extended to further Y-chromosomal extraction loci, whereas different probe sets for each extraction locus could be designed and analyzed. Hereby, one probe set defines a number of probes with different lengths targeting the same extraction locus with the same orientation and the same 3'-terminal mismatch. For example, the extraction probe P224FC already described in chapter 3.4.1 presents one tested

probe set which comprises 14 different probes of the length 23 and 21 to 9 nucleotides. Furthermore, the advanced primer design program Visual Omp™ was used for the exact evaluation of all probe properties. Visual Omp™ allows next to the determination of typical primer properties such as primer length, GC contents, melting temperature (Tm), GC clamp, dimers (including cross-dimers and self-dimers), hairpin structure and specificity, also the calculation of precise thermodynamic values for ΔS, ΔH, ΔG, and duplex concentrations for each occurring primer template dimer dependent on temperature.

The effect of different probe lengths in HSE reaction was tested at eight additional extraction loci: P30, P38, P240, P244, Tat, rs13304202, S4:2701 and S6:4204 (chapter 3.1). In total tested 41 probe sets were tested, which combine the analysis of 137 different HSE probes. The probe sets showed a similar dependency between HSE success and probe length as the previous described at the probe set P224FC and FT (Figure 3-8).

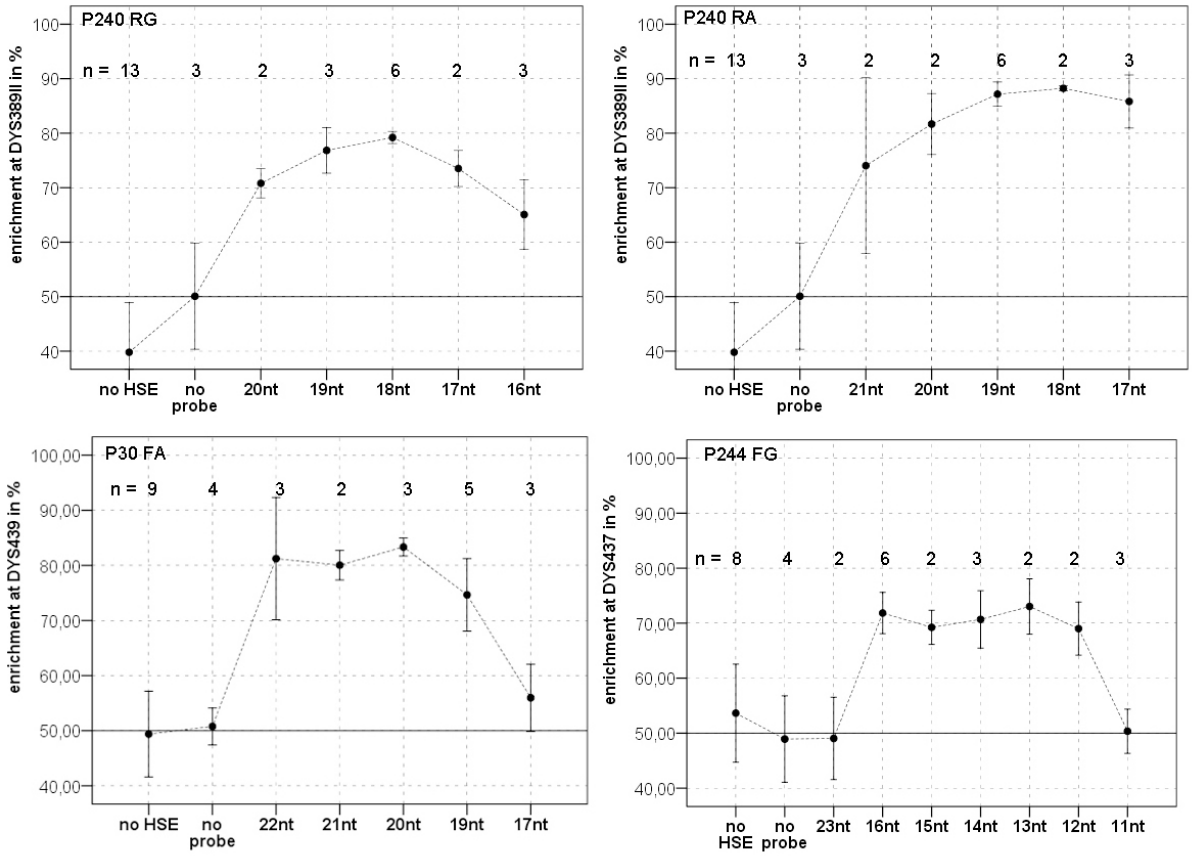


Figure 3-8: Separation success of HSE reactions depending on different probe lengths.
Enrichment of one contributor from a male DNA mixture is shown in relation to probe length for four further tested different probe sets P240RG/RA, P30FA, and P244FG. HSE was carried out with male DNA mix 1114/94447, 5573/5574 and 5573/95210. The bars indicate standard deviation of the mean enrichment. "No HSE" indicates analysis of the male DNA mixture without separation by HSE, and "no probe" indicates separation of a male DNA mixture by HSE without probe, n = number of extractions.

Table 3-3 provides a summary of the 41 probe sets tested in 345 HSE reactions with their corresponding G/C content and the alignment results for each probe (Table 3-3A). Furthermore the table shows the affiliation of the probe sets to their nine extraction loci and the distance to the closest STR marker. The extraction sites M170, M173, M198 and M343 are not presented in Table 3-3, because they did not show any HSE success. The given probe lengths correlate to the best length with highest HSE success (Appendix-Table 7-3) and confirmed the presence of an optimum probe length for HSE, that varied for each targeted locus. Furthermore the probes were analyzed after their match and mismatch hybridization which is summarized in the columns B to D in Table 3-3. The column B shows the differences of the ΔG between match and mismatch as a ΔΔG (described in detail further below). Colum C summarizes the simulated best probe length derived from the comparison of the heteroduplex concentrations between match and mismatch and column D shows the mach and mismatch base for each probe set (described in chapter 3.4.3).

The variation of the maximum enrichment of one contributor generally depends on the different distance of each extraction locus from the STR marker and the sequence specificity of the probe. For example, the probe S6:4204 showed a second target site with high similarity located very close to the intended target locus (Figure 3-9). This unspecific match had one internal mismatch and a mismatch of the 3'-termini only for probe S6:4204 FC. The 3'-terminus of the probe S6:4204 FG matched completely at this secondary locus, which resulted in a significant reduction of the separation effect at the intended target locus from 98% to 67% enrichment.

Table 3-3: Summary of HSE success and simulated parameters for all tested probe-sets.

A						B	C		D	
probe - distance to STR - Marker		mean GC%	distance of hits in kb with matching 3'end [1]	HSE results [3]		ΔΔG (M-MM)	simulation with viusal omp		match (target-probe)	mismatch (target-probe)
				success in %	best tested length		best length	HSE fits with simulation		
P30 - 18kb	FG	36	no	**76**	21	-1.15	21, 22	r	C-G	T-G
	FG-1	35	no	**82**	21	-1.8	21	r	C-G	T-G
	FA	33	no	**83**	20	-0.4	21	r	T-A	C-A
	FA-1	30	no	**95**	22	-1.01	22	r	T-A	C-A
	RC	27	no	**92**	24	-0.49	24	r	G-C	A-C
	RC-1	30	2, 24	**70**	23	-2.18	23, 27	r	G-C	A-C
P38 - 17kb	FA	77	no[2]	**89**	21	-0.6	11	w	T-A	A-G
	FA-1	71	no[2]	**86**	20	-1.04	12	w	T-A	A-G
	FC	83	2, 9, 10, 13 +	**72**	15	-1.19	10	w	G-C	T-C
	FC-1	79	7, 10 +	**49**	n.s.	-1.83	12	w	G-C	T-C
P224 - 13kb	FC	40	no	**90**	19	-0.37	18	r	G-C	A-C
	FT	34	no	**86**	19	-0.7	19	r	A-T	G-T
	RG	60	no	**93**	17	-1.31	14	r	C-G	T-G
	RA	53	no	**86**	17	-0.48	15, 16	r	T-A	C-A
P240 - 14kb	FC	37	no	**82**	22	-0.48	22	r	G-C	A-C
	FC-1	40	no	**83**	20	-2.43	20, 21	r	G-C	A-C
	FT	32	no	**84**	22	-0.6	22	r	A-T	T-C
	FT-1	35	no	**94**	20	-1.52	20, 21	r	A-T	T-C
	RG	42	no	**79**	18	-1.14	19	r	C-G	T-G
	RG-1	40	no	**73**	20	-1.21	20	r	C-G	T-G
	RA	39	no	**88**	18	-0.87	18, 20	r	T-A	C-A
P244 - 35kb	FG	66	49	**73**	13	-0.95	15	r	C-G	T-G
	FG-1	65	no	**78**	15	-2.11	14	r	C-G	T-G
	FA	57	116 +	**54**	n.s.	0.05	no	r	T-A	C-A
	FA-1	59	0,6 +	**76**	15	-1.04	15	r	T-A	C-A
	RC	68	20, 29 +	**68**	14	-0.66	13	r	G-C	A-C
	RC-1	71	20 +	**71**	14	-2.16	14	r	G-C	A-C
	RT	60	2, 9, 13 +	**62**	13	-0.69	13	r	A-T	G-T
	RT-1	64	9, 13 +	**66**	14	-1.01	14	r	A-T	G-T
Tat 15kb	FT	33	no	**84**	20	-0.57	20	r	T-A	G-T
	FC	38	no	**87**	18	-0.52	20	r	G-C	A-C
rs13304202 - 37kb	FG	40	no	**75**	19	-1.39	20	r	C-G	T-G
	FG-1	43	no	**81**	18	-2.77	19, 20	r	C-G	T-G
	FA	34	no	**77**	19	0.42	no	w	T-A	C-A
	FA-1	38	no	**76**	22	-1.55	20, 22	r	T-A	C-A
	RC	39	no	**70**	21	-1.17	22	r	G-C	A-C
	RC-1	43	no	**81**	21	-3.00	21, 24	r	G-C	A-C
S4:2701 1.3kb	RG	35	no	**78**	20	-1.19	19	r	C-G	T-G
	RA	30	no	**91**	20	-0.13	20	r	T-A	C-A
S5:4204 0,15kb	FG	53	0.3kb with matched 3' end	**67**	17	-0.71	18	r	C-G	G-G
	FC	53	0.3kb but no matched 3' end	**98**	17	-0.83	14	r	C-G	C-C

(A) and (B) Comparison of HSE results for all probe sets with distance to the STR marker, G/C content, hit frequency and free energy differences of match and mismatch in kcal/mol. (1) The number is the distance in kb of strong hits to the extraction locus after BLAST, + indicates the presence of further hits but greater distance. (2) Best simulated probes P38FA (11 nucleotides long) and P38FA-1 (12 nucleotides long) showed many hits (4, 12, 16 kb+ and 4, 10, 16 kb+). (3) HSE results show the best experimentally evaluated probe length with enrichment of one contributor. Low separation in HSE could be due to a higher hit frequency close to the extraction locus and great distance to the STR marker. (C) The best probe length was evaluated from simulation and chosen after highest Δcon and comparison of the prediction for best probe length after Δcon and best probe in HSE: (r) right prediction, (w) wrong prediction. (D) Listing of all analyzed match and mismatch constellations in this study. List of complete data is available in Appendix-Table 7-3.

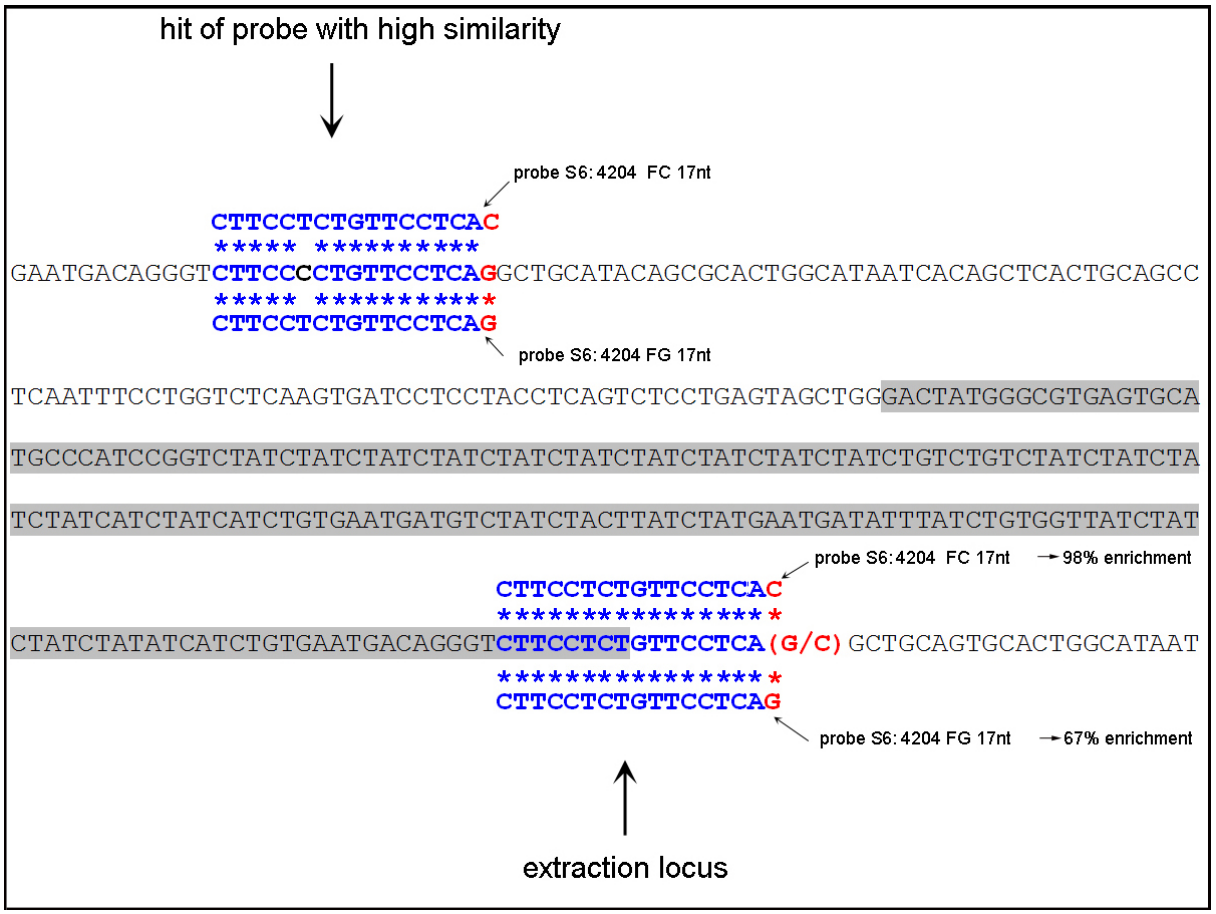


Figure 3-9: Extraction locus S6:4204.
The probe S6:4204 shows 284 bp upstream of its extraction locus a second hit with high similarity sequence. All blue asterisks indicate same sequence and red asterisks indicate match of the 3'-termini. Grey box marks DYS437.

For an initial analysis of the results, the HSE separation success was compared for all investigated probes in relation to probe length, free energy, melting temperature and G/C content (Appendix-Table 7-3). Table 3-4A provides these results for the P224FC and P30RC probe sets. As probe length increased, melting temperature and free energy also increased, though probe G/C content is a function of sequence identity. Optimum separation was obtained using probes 15-20 nucleotides in length that ranged in melting temperature from 54-59°C, had a free energy of -6.56 to -8.69 kcal/mol and a G/C content of 37 to 47%. Comparatively, for the P30RC probe set, optimum separation was obtained with probes 24 or 25 nucleotides in length with a melting temperature of 57°C, a free energy of -7.63 to -7.55 kcal/mol and a G/C content of 28 to 29%. Since the GC contents and melting temperature are a consequence of the probe sequence they vary between the different probe and show different values for probes with best HSE success. For example the probe set P240 FC-1 and P30 show similar GC contents (35% and 40%) but best separation at a melting temperature of 60 or 56°C, respectively. On the other hand also probes with very little (P30 RC, 29%) or very high (P240 FC-1, 60%) GC contents could achieve best HSE success (Appendix-Table 7-3). These data

demonstrate that the differences in target sequence complexity from one target to the next must obviously influence the optimum design attributes for each probe, even though these attributes are not by themselves reliable indicators of HSE efficiency. Also, the comparison of all probes according to their separation effect did not reveal any significant optimal parameters.

Table 3-4: Probe design for P224 FC and P30RC.

A

probe - STR-Marker	l	probe sequence	GC %	Tm_M	$\Delta G°_M$	HSE result in %	n
P224FC - DYS390	25	TTTCAGAAATGAGTGTGACATCTTC	36	63	-10.97	-	
	24	TTCAGAAATGAGTGTGACATCTTC	38	63	-11.09	-	
	23	TCAGAAATGAGTGTGACATCTTC	39	63	-10.73	**68 ± 11**	4
	22	AGAAATGAGTGTGACATCTTC	41	63	-10.8	-	
	21	GAAATGAGTGTGACATCTTC	38	61	-9.58	**85 ± 4**	3
	20	AAATGAGTGTGACATCTTC	40	59	-8.69	**89 ± 5**	6
	19	AATGAGTGTGACATCTTC	37	59	-8.32	**91 ± 5**	10
	18	ATGAGTGTGACATCTTC	39	58	-7.95	**89 ± 7**	7
	17	TGAGTGTGACATCTTC	41	57	-7.58	**88 ± 9**	8
	16	GAGTGTGACATCTTC	44	54	-6.4	**91 ± 7**	7
	15	AGTGTGACATCTTC	47	54	-6.56	**92 ± 4**	6
	14	GTGTGACATCTTC	43	52	-6.02	**86 ± 6**	6
	13	TGTGACATCTTC	46	49	-5.15	**84 ± 6**	5
	12	GTGACATCTTC	42	45	-4.26	**70 ± 5**	3
	11	TGACATCTTC	45	44	-4.01	**79 ± 5**	3
	10	GACATCTTC	40	36	-2.85	**57 ± 6**	3
	9	GACATCTTC	44	35	-2.6	**54 ± 7**	3
	8	ACATCTTC	38	26	-2.06	-	
	6	ATCTTC	43	21	-0.93	-	
P30RC - DYS439	26	TTCTATCCATCTATCATCTATTTATC	27	**57**	-7.43		
	25	TCTATCCATCTATCATCTATTTATC	28	**57**	-7.55	**86 ± 1**	2
	24	CTATCCATCTATCATCTATTTATC	29	**57**	-7.63	**92 ± 9**	2
	23	TATCCATCTATCATCTATTTATC	26	**54**	-6.02	-	
	22	ATCCATCTATCATCTATTTATC	27	**55**	-6.62	**51 ± 1**	2
	21	TCCATCTATCATCTATTTATC	28	**53**	-5.47	-	
	20	CCATCTATCATCTATTTATC	30	**54**	-6,04	-	
	19	CATCTATCATCTATTTATC	26	**49**	-4.14	**46 ± 8**	**2**
	18	ATCTATCATCTATTTATC	22	**47**	-3.56		

B

con_M *10-^{17}M	con_{MM} *10^{-17}M	Δcon_{M-MM} *10^{-17}M
297	290	**7**
297	292	**6**
295	285	**10**
296	287	**9**
274	233	**41**
219	141	**79**
182	100	**82**
141	68	**73**
100	43	**58**
23	9	**14**
29	12	**17**
13	5	**8**
4	1	**2**
1	0	**1**
1	0	**0**
0	0	**0**
0	0	**0**
0	0	**0**
86	48	**38**
97	56	**42**
106	61	**44**
13	7	**7**
31	16	**16**
6	3	**3**
14	7	**7**
1	0	**0**
0	0	**0**

Probe P224FC and P30RC were designed for SNP P224 (C) and SNP P30 (G) as the extraction loci in forward (F) and reverse (R) orientation. (A) For different probe lengths (l), the G/C content, melting temperature (Tm), and ΔG value were calculated using Visual Omp™. The success of separation by HSE with different probe lengths is given as the percent enrichment of one contributor with standard deviation and number of experiments (n). (-) indicates no data, shades of grey represent separation success: dark grey indicates that the enrichment of one contributor is > 88% and the contributing allele appears only as a stutter, grey indicates enrichment between 61% and 88% and light grey indicates separation < 61%. (B) Concentrations of hetero-duplex target probes (con) for match (M) and mismatch (MM) have been simulated for different probe lengths using Visual Omp™ with the assay parameters. Δ con presents the difference between match and mismatch concentrations.

Next, match and mismatch hybridization have been compared by calculating $\Delta\Delta G_{M-MM}$. Within one probe set, ΔG decreased slowly with increasing probe length (Figure 3-11 C) but did not exhibit any consistency with regard to probe-specificity in HSE. Comparison of the ΔΔG values between different probe sets did not reveal a correlation between either the extent of ΔΔG and HSE success or the type of mismatch. For example, the probes rs1330202RC and P240FC both had a G-C mismatch, but their

ΔΔGs were -1.17 and -0.48 kcal/mol, respectively, which are rather different free energy contributions. The probes P224 FC and RG had very different ΔΔG values as well, -0.37 and -1.31 kcal/mol, respectively, but both had very high enrichment (>90%) of one contributor (Table 3-3B).

3.4.3 Prediction of probe specificity after differences in match vs. mismatch concentrations

Since so far all comparisons of several probe parameters could not explain the observed correlation between HSE success and probe length, also the simulated match and mismatch concentration have been compared. Hereby, it was observed that the simulated match and mismatch concentrations naturally increased with probe length, but their concentration difference ($\Delta con_{M\text{-}MM}$) exhibited an optimum curve that is dependent on probe length (Figure 3-11A und D). For example, the simulation for P224FC showed that, with increasing probe length, $\Delta con_{M\text{-}MM}$ first increased and then decreased dramatically when probes were longer than 21 nucleotides (Table 3-4B). The properties of match and mismatch concentrations were also examined in melting-curve analyses (Figure 3-10). Naturally, the duplex concentrations for both matches and mismatches decrease with increasing temperatures for all tested probes. Two general observations can be made: First, the melting curve of the mismatch duplex shows smaller concentrations than the matched hybridizated probe. Second, the simulated melting curves are shifted to higher temperatures with increasing probe lengths.

When examining duplex concentrations at the hybridization temperature, melting curve diagrams illustrate very clear the change of Δcon with increasing probe length. Figure 3-10 exemplifies this fact on the probe set P224 FC, which is also described in Table 3-4. HSE with P224 FC (23 nucleotides long) showed a separation effect with a ratio of only 2:1. The simulated duplex concentrations for probe P224FC are very high for both, match and mismatch, and therefore Δcon remains small. For the same probe however only 19 nucleotides long, the difference between concentration of match and mismatch probe increased, but decreased again when duplex concentration dropped down. Tracking Δcon, HSE success also first increased up to a ratio of 50:1 for the 19 nucleotides long probe but then again decreased to a ratio of only 2:1 for probe P224 FC, 12 nucleotides long. For all other tested probe sets except for P244FA and rs13304202FA, a similar increase and decrease of Δcon with longer probes could be observed.

The observation of an optimum Δcon curve reflects the separation effect of the tested probes fairly well. For example, the simulation of Δcon for probe P224 FC showed a high Δcon values for a range of probe lengths from 17 to 21 nucleotides. This is in agreement with the HSE success observed, which occurred also for broad optimum from 20 to 15 nucleotides (Figure 3-7 A, Table 3-4). For

probe P30 GA, optimum Δcon occurred only for a shorter range of 20 and 22 nucleotides, which completely agrees with the observed separation maximum for HSE (Figure 3-8, Appendix-Table 7-3). Thus, single Δcon values do not provide information about probe specificity but have to be regarded as relative values within one probe set to determine the best probe length for the maximal concentration difference between matched and mismatched probes. For example, the maximum Δcon of different probe sets can have very different values but still exhibit the same probe specificity (Figure 3-11D).

For the nine selected extraction loci, over 500 probes were simulated and compared to the separation success of 345 HSEs. In total 88% (= 36 probe sets) of the analyzed data showed accordance of the simulated Δcon curve with the obtained HSE results (Table 3-3).

However, five probe sets (12% of all tested probe sets), which refer to probe P38 and rs13304202 FA, did not show any accordance between Δcon and the HSE success (Table 3-3). The maximal simulated Δcon for probe P38 occurred with 11 or 12 nucleotides, but in HSE the optimum was found to be at a longer probe length (>15 nucleotides, Table 3-3). For rs13304202FA, Visual Omp™ predicted that mismatches were more stable than their matches. Therefore, no separation by HSE was predicted, though some separation was observed with the best length of 19 nucleotides. In the case of P244FA, matched and mismatched targets were observed in which both free energies were predicted to be equal and, as expected, no specific extension was observed.

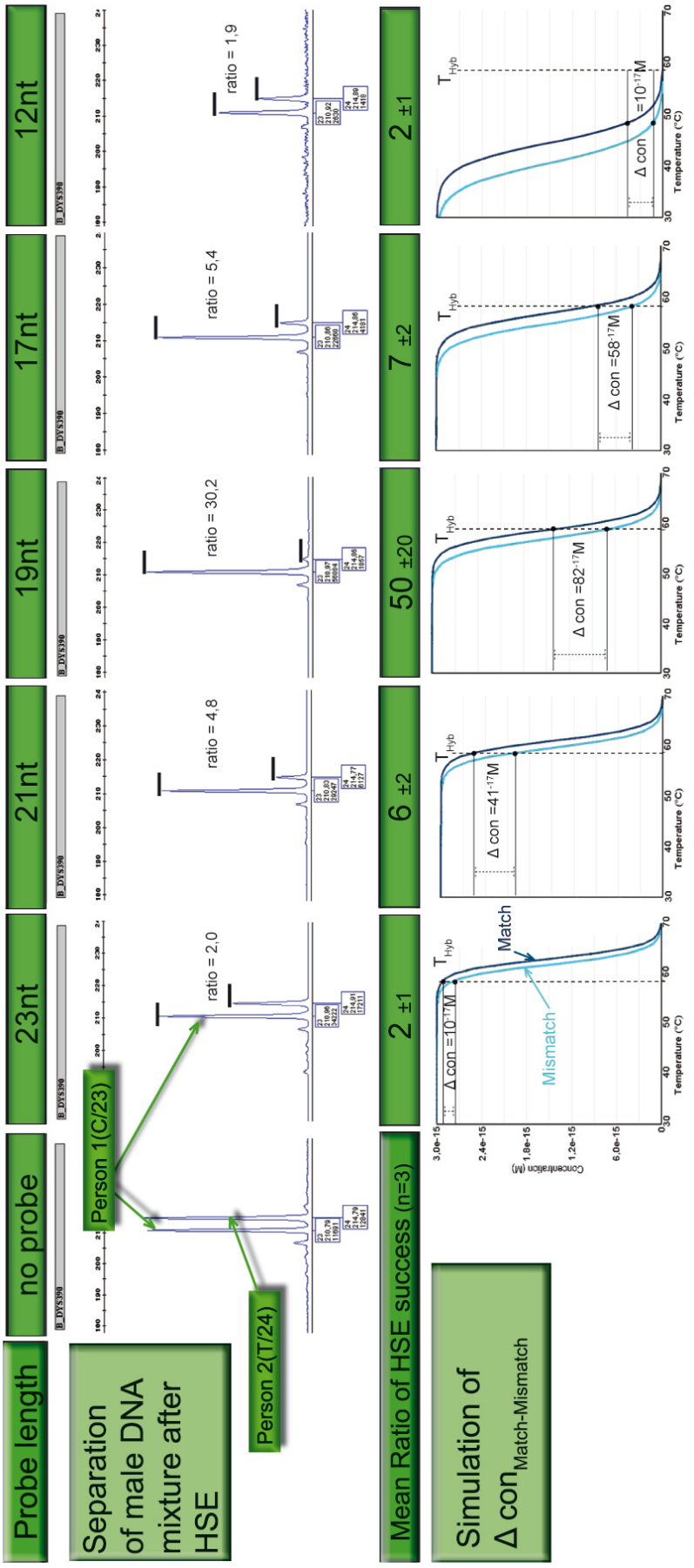


Figure 3-10: Comparison of HSE success and simulated Δcon values for different probe lengths.
Upper panel shows ratio between the two male DNAs of the mixture (person 1 = 5573, person 2 = 1118) for locus DYS390 after HSE without probe or with different lengths of probe P224FC. For the same probes simulated Δcon values at hybridization temperature (T_{Hyb}) are shown in the lower panel and illustrated as concentration versus temperature diagram (see chapter 2.13).

Analysis of mismatch positions, mismatch types, and G/C content

For probe sets with a low separation effect, probes were also tested with an internal mismatch next at the penultimate (-1) position of the 3'-termini. There is indication from the literature that mismatches at the -1 position may be more disruptive to the polymerase and thus more effective in allelic discrimination for the purpose of HSE. In this study, 15 different probe sets with internal mismatches were tested and could be compared to their associated probe sets with terminal mismatches (Table 3-3A). An increased enrichment of over 10% was observed for only three probe sets (rs13304202RC, P240FT, and P30FA) when the mismatch was located at the penultimate position. In contrast, probe sets P30RC-1 and P38FC-1 exhibited a decrease in enrichment of 22% and 23%, respectively. Furthermore, in this study HSE were carried out with probes showing eight different types of mismatches; the mismatches (target-probe) A-C, C-A, and T-G occurred with the highest frequency (Table 3-3D). A comparison of seven A-C/C-A mismatches contra G-T/T-G mismatches did not reveal any significant effect on HSE success for one mismatch. However, more comprehensive data are necessary for better evidence of the influence of type and position of mismatches for specific enzymatic extension reactions.

The target hybridization efficiency of six core probe match and mismatch sets was simulated in Visual Omp™. Each of the core probe sets vary in G/C content from 32% to 77% and were analyzed for the lengths from 6 to 46 base pairs (Figure 3-11). Each probe length was plotted as an independent variable against the dependent variables of probe concentration (con) and duplex free energy (ΔG) (Figure 3-11A and C). A clear dependence was observed between probe length, G/C content, and concentration, as well as for ΔG, as expected, though no correlation with HSE efficiency was observed. Next, the changes in ΔΔG and Δcon were compared among the six core probe match and mismatch sets based on probe length. For the plot of ΔΔG as a function of probe length, no correlation with GC content or probe length was observed (Figure 3-11B). As described already before the observed HSE success of the different probe sets could not be correlated to their ΔΔG (chapter 3.4.2).

Δcon is presented as a function of probe length in Figure 3-11D. Here, Δcon appears as a parabolic curve in which the maximum Δcon shifts to longer or shorter probe lengths depending on the G/C content. The amplitude of the maximum or shape of the curve is different between all probe sets and does not show any dependence on G/C content. Because of the partial dependency of Δcon on G/C content, the best predicted probes from all 41 tested probe sets were compared based on G/C content and probe length. Figure 3-11E shows that the probe length at which Δcon reached its maximum decreases with increasing G/C content.

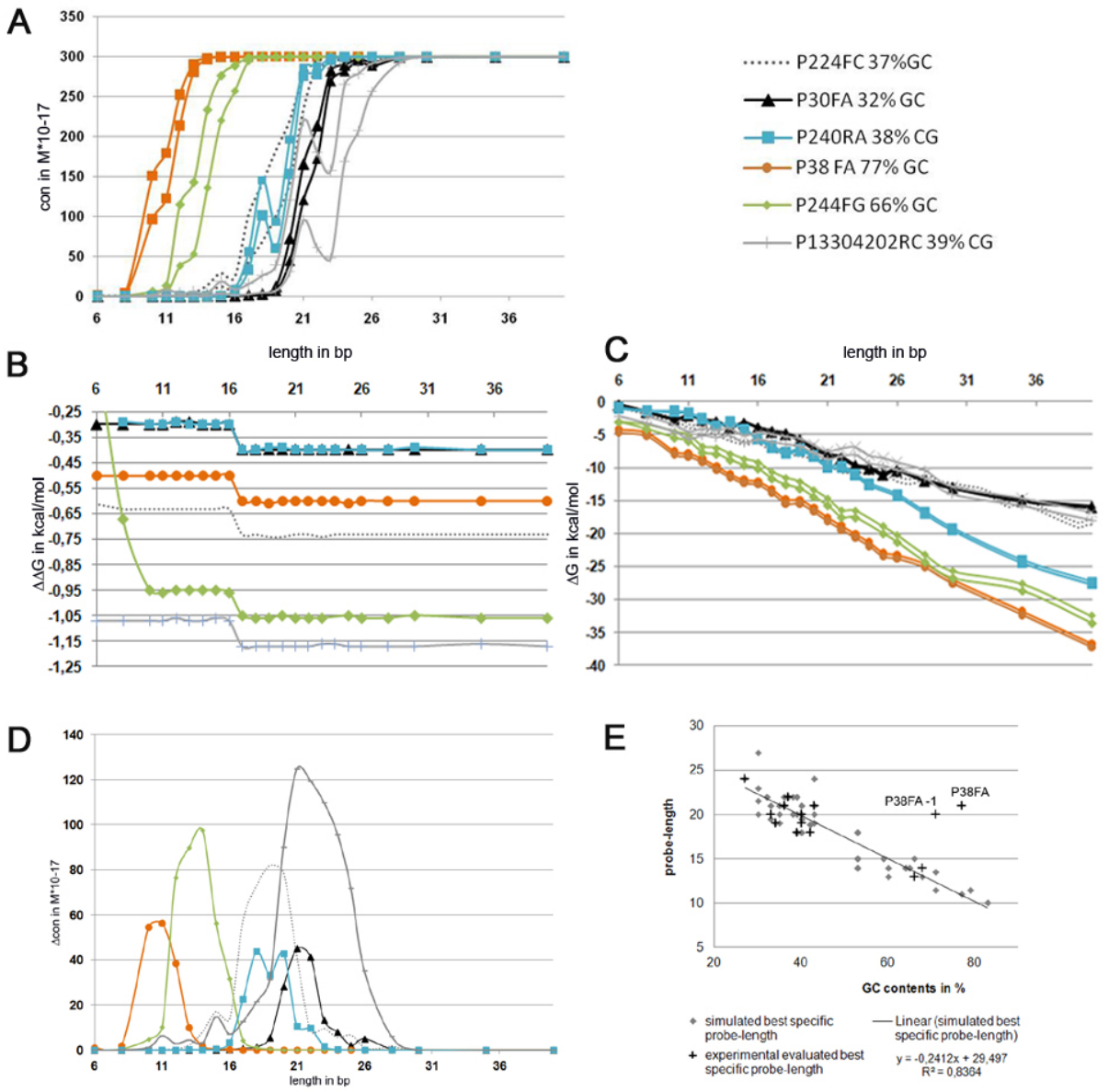


Figure 3-11: Comparison of probes with different GC contents in dependence of their lengths.
The GC content was calculated as mean values for all simulated probe lengths of one probe set. (A) Concentration of target-probe heteroduplexes for match and mismatch in dependence of their lengths and GC contents. (B) ΔΔG of target-probe heteroduplexes for match and mismatch in dependence of their lengths and GC contents[4]. (C) ΔG decreases with increasing probe length and GC content. (D) Comparison of Δcon with different G/C content (E) Diagram showing the linear dependence of probe length and G/C content for best specific probe lengths based on simulation versus actual HSE extraction experiments. Data points are based on Table 3-3.

When the G/C content was plotted against the probe length of the most specific probes evaluated in HSE experiments (i.e. independently of the simulation Figure 3-11E, crosses), a nearly linear

[4] The presence of the "nick"-shaped artifact that is visible in the thermodynamics plot that occurs from 16 to 17 nucleotides was discussed with experts from Visual OMP (Dr. Norm Watkins). They indicated that it was most likely caused by a polymer salt correction, where the condensation of salt ions changes at this nucleotide threshold when calculating the polymer salt effect in reference to Manning's electrolyte condensation theory [48, 56] .

dependency could be observed. Using linear regression (formula in Figure 3-11E), it was possible to predict optimal probe length by G/C content, which matched the simulated and experimentally evaluated optimal probe length within only two nucleotides (data not shown).

3.4.4 Required proximity of extraction locus to detected marker

One basic question of HSE design is how the distance between the extraction site and the analyzed marker influences the HSE success and what is the maximum length of the separated fragments. Therefore twelve of in this study used extraction sites where compared after their distance to the nearest Y-filer STR markers and their best obtained HSE success (Table 3-5). Furthermore, some selected extraction loci were located also close to a second Y-filer STR marker and therefore HSE success of the second marker could be also compared with a second distance (Figure 3-15). For example, the extraction probe P240 is located in 14 kb distance to the nearest STR marker DYS389 and in 83 kb distance to the second closed Y-filer marker DYS439. The comparison of all twelve extraction loci and their nearest Y-filer marker shows that HSE reactions with a distance of less than 1 kb between target and detection site showed a very high separation success with over 90% enrichment. HSE with distances between 11 kb and 19 kb also resulted in a significant or complete HSE separation (85% to 95% enrichment), whereas extractions with distances around 35 kb showed a decrease of the separation effect to 81% and 76% enrichment (Table 3-5).

The comparison of the five second nearest Y-filer marker showed a separation effect with an enrichment over 60% for the extraction sites P30, P244 and rs13304202 in 30 kb, 53 kb and 51 kb distance, but with a high variation for the extraction site P30. Therefore second nearest markers were analyzed for a higher number of successful HSE reactions including all probe sets of one extraction site. For example, HSE reactions with probes for the extraction site P30 showed in 37 reactions an enrichment > 70% for the nearest Y-filer marker DYS439 and were analyzed for the enrichment of the second nearest marker DYS437 with an mean enrichment of 66% ±9 with an significance of <0,001 (Mann-Withney test). Similar results were also obtained for the second nearest marker of the extraction sites P244 and rs13304202 (66% ±9, 66% ±6, n=18). No enrichment was detected for the second distant marker in HSE with P38 (57% ±8, n=9) and P240 (54% ±5, n=28).

When plotting the mean enrichments according to their distances (Figure 3-12), the separation effect decreases constantly with rising distances between the extraction locus and the analyzed Y-STR markers. Hereby, the results of the nearest and the second nearest Y-filer marker show a linear decline, except for the second nearest marker of P30 and especially of P38 (Figure 3-12). The maximum observed distance at which a separation effect was still observed is at about 50 kb.

Table 3-5: Summary of distances between HSE extraction loci and analyzed Y-STR marker.

best extraction probe	nearest Yfiler STR locus		second nearest Yfiler STR locus		
	distance	mean enrichment[1]	distance	mean enrichment[1]	n
S6:4204 FC 17nt	0,15kb-DYS437	98%±1	n.a.		2
S1:2701 RA 20nt	1,3kb-DYS437	91%	n.a.		1
P224 FC 19nt	11kb-DYS390	91%±5	n.a.		10
P240 FT-1 20nt	14kb-DYS389	94%±5	83kb-DYS439	53±5	3
Tat FC 18nt	15kb-DYS438	87%±1	n.a.		3
P38 FA 19nt	17kb-DYS437	85%±6	31kb-DYS439	59±12	2
P30 FA-1 22nt	19kb-DYS439	95%±2	30kb-DYS437	63±18	2
P244 FA-1 15nt	34kb-DYS437	76%±4	53kb-DYS635	61±2	3
rs13304202 RC-1 21nt	37kb-DYS635	81%±8	51kb-DYS437	68±5	2
M173	88kb-DYS438	49%±19	n.a.		8
M170	90kb-DYS438	49%±12	n.a.		4
M198	93kb-DYS438	51%±17	n.a.		4

(1) Enrichment is given as the mean value with standard derivation for nearest and second nearest Y-filer STR loci. Probe was selected after best observed HSE success at the respective extraction locus. (n) is number of experiments. Values are taken from Appendix-Table 7-3.

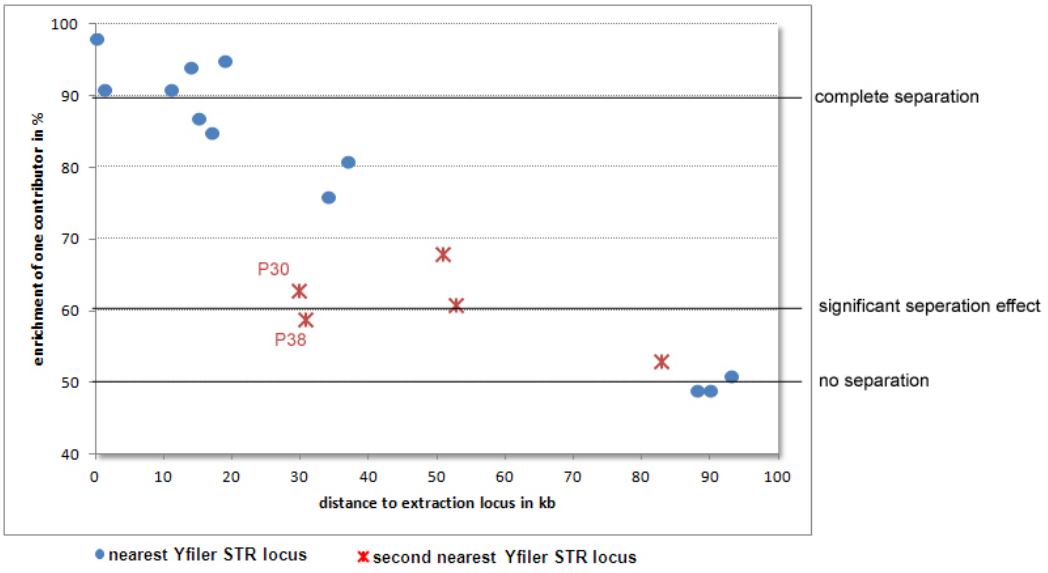


Figure 3-12: HSE success in dependence of the distance to the extraction locus. Diagram summarizes the results from Table 3-5.

3.4.5 Confirmation of the new prediction model for probe design at a further locus in the genome

A reliable correlation between probe specificity and the simulated concentration differences of matched and mismatched probes ($\Delta con_{M\text{-}MM}$) has been observed and evaluated for the separation of STR markers on the Y-chromosome. For further validation, the new probe design was applied and tested on the entire different HLA locus on chromosome 6, which exhibits an exceptionally high degree of variation of complex haplotypes.

As a part of a parallel project in our laboratory, HSE probes were designed for the determination of *Macaca mulatta* (Rhesus macaque) MHC-DQB1 alleles. For this project, HSE probes were used to isolate clusters of Mamu MHC-DQB1 haplotypes with similar characteristics in order to simplify data interpretation, especially in the case of ambiguous alleles. The first probe design studies were carried out on the DQB1 locus, because of its less complex structure. Probes were tested first with human DNA since the ape DNA resources were limited. This was possible because the MHC loci of human and ape show high similarities and the chosen extraction locus is conserved in both species.

The probes DQB1:05-06 and DQB1:02-04 were designed for two point mutations (A/CGGA/C, see Figure 3-13) at the end of exon 2, which separate the two clusters HLA-DQB1*05 to*06 and DQB1*02 to *04 in human or respectively the MHC-DQB1*06 and MHC-DQB1*14 to *24 clusters in the macaque. The designed probes anneal completely to the reference sequence HLA-DQB1*05:01:01:01 in reverse orientation, with the 5'-termini in the adjacent intron region. Therefore, for some alleles the designed probes could show some mismatch at the 5'-end. For example, in this study the used human DNA sample exhibited the alleles, HLA-DQB1*06:04, for which the intron sequence was not available, and HLA-DQB1*02:01 which show one mismatch at the 5'-end of the probe.

Consistent with the probe design studies on the Y chromosome (chapter 3.4.2), HSE reactions and simulations of duplex concentrations were carried out for different probe lengths. Probe DQB1:05-06 was tested for the lengths 22, 18 and 16 nucleotides, and probe DQB1:02-04 for the lengths 22, 18 and 14 nucleotides. The tested probe-lengths were selected after poor, average and best simulated $\Delta con_{M\text{-}MM}$ values. Probe DQB1:05-06 showed the best concentration difference between match and mismatch for the 15 and 16 nt long probes, and probe DQB1:02-04 for 13-14 nt (Table 3-6).

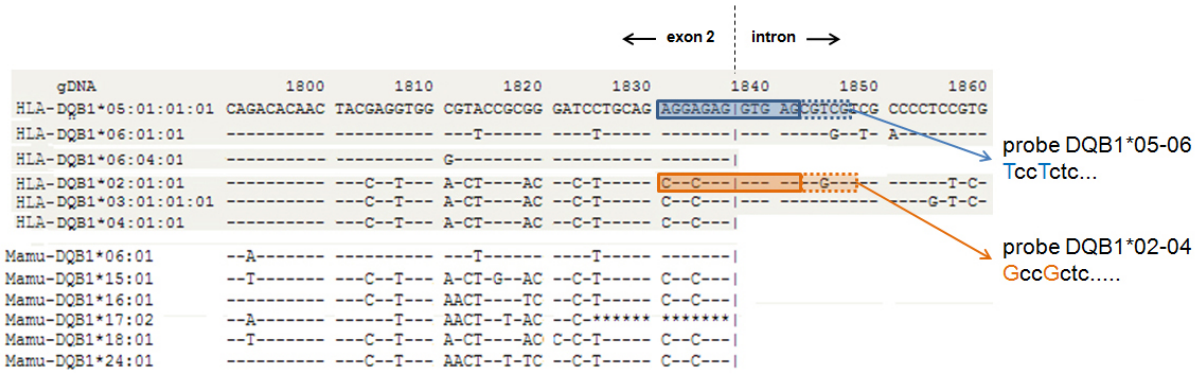


Figure 3-13: Probe design for HSE-based separation of rhesus Mamu DQB1 haplogroup cluster. Sequence data and positions refer to numbered full length alignment with the IPD IMGT/HLA database.

Table 3-6: Simulation of Δcon for HSE probes at the DQB1 locus.

probe DQB1:05-06				probe DQB1:02-04			
probe length	con_{Match} in M^{-17}	$con_{Mismatch}$ in M^{-17}	Δcon_{M-MM} in M^{-17}	probe length	con_{Match} in M^{-17}	$con_{Mismatch}$ in M^{-17}	Δcon_{M-MM} in M^{-17}
23	300	300	0	22	300	300	0
22	300	297	3	20	300	299	1
20	300	297	3	19	300	291	9
19	300	257	43	18	300	291	9
18	300	264	35	17	300	257	43
17	299	171	127	16	299	102	197
16	281	31	251	15	299	117	182
15	285	37	248	14	297	31	266
14	231	7	224	13	269	3	266
13	67	1	66	12	195	1	194
12	17	0	17	11	80	0	80

Grey bars show probe length tested in HSE (see Figure 3-14).

In contrast to the previously described separation of Y-chromosomal STR-markers, HSE samples were analyzed by HLA sequencing and not by the Y-filer multiplex PCR. Due to this, the results are shown as sequencing files and diploid genotype can be seen by two different colored peaks which represent two different bases at the same position. After HSE reactions with the DQB1-probes, these mixed positions showed an enrichment of the bases corresponding to the targeted haplotype (Figure 3-14). Hereby it was found that the HSE success increase with increasing Δcon values and that best probes length correlated with highest Δcon values. For example, at the reference position 1691, the DQB1 HLA allele *06:04 has a thymine and the allele *02:01 has an adenine, which shows equal peak intensities in the non-separated sample (Figure 3-14: grey part). When haplotype *02:01 was separated by HSE with probe DQB1:02-04 22 nucleotides long, the peak intensity of the adenine and therefore the HSE of the DQB1*02:01 allele rose to a 2-fold ratio. When HSE was carried out with the shorter probes 18 and 16 nucleotides long, the ratio increase to 2.5 and 5-fold. Hereby, the probe

DQB1:02-04 with the highest simulated Δcon value (16nt long) also showed the highest observed separation effect (Figure 3-14 blue part). Conversely, the HSE reaction with the probe DQB1:05-06, 16 nt long, resulted in an enrichment of the thymine base, corresponding to the DQB1*06:04 allele with almost complete separation for DQB1*06:04 (Figure 3-14 yellow part).

HSE success obtained clearly depends on probe length, and the best separation effects were observed by the probe with largest simulated Δcon value. The best probes according to this study (DQB1:02-04 14nt and DQB1:05-06 16nt) were successfully used for the separation of ambiguous DQB1 alleles of the Rhesus macaque (data not shown).

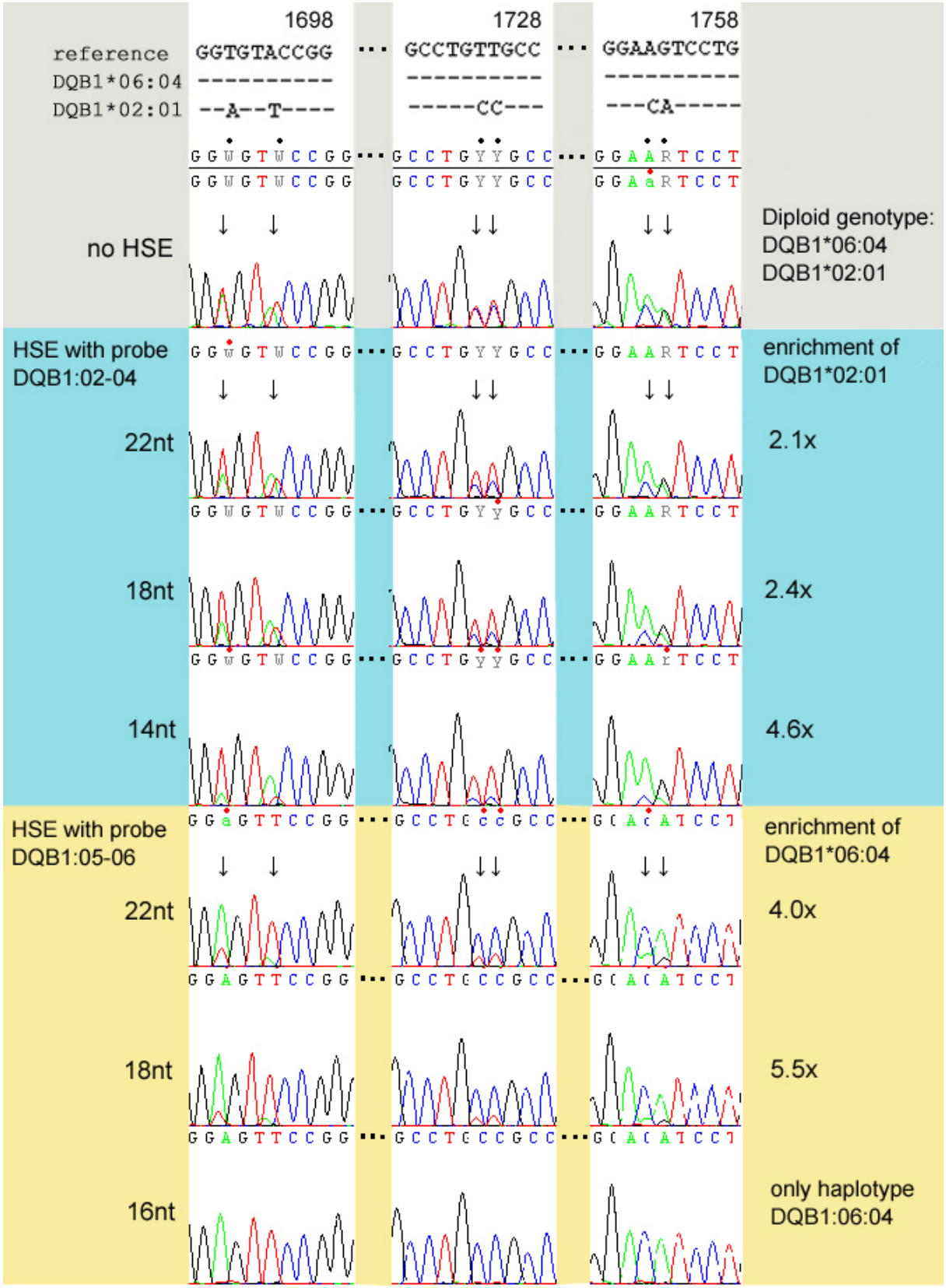


Figure 3-14: Separation of the human DQB1 locus by HSE with probes of different lengths.
Arrows indicate heterozygous position. Lengths of probes are given in nucleotides (nt). Positions refer to numbered full length alignment with the IPD IMGT/HLA database.

3.5 First studies of HSE for forensic applications

The application of HSE for the separation of forensic mixtures requires an effective enrichment of individual markers as well as the investigation of additional aspects, such as the quality of forensic samples and the separation of entire individual Y-chromosomal haplotypes. In this study a first HSE multiplex approach could be established, which combines several Y-STR markers. Furthermore HSE was tested with different amounts of DNA mixture.

3.5.1 Design of Y-chromosomal multiplex HSE

Y-chromosomal haplotypes carry more information about individual identification when more than one STR-marker can be separated. Therefore, a HSE reaction was designed which multiplexes five probes or in other words uses five extraction sites (rs13304202, P30, P224, P240 and either P38 or P244) in one separation reaction. Selected target loci were located next to the markers, DYS635, DYS437, DYS439, DYS389, and DYS390 with a minimal distance of 11 kb and a maximum distance of 37 kb between the extraction site and the corresponding DYS marker (Figure 3-15). Hereby, five of the extraction sites were located between a cluster of DYS markers and therefore some of them showed relatively short distances to two DYS markers.

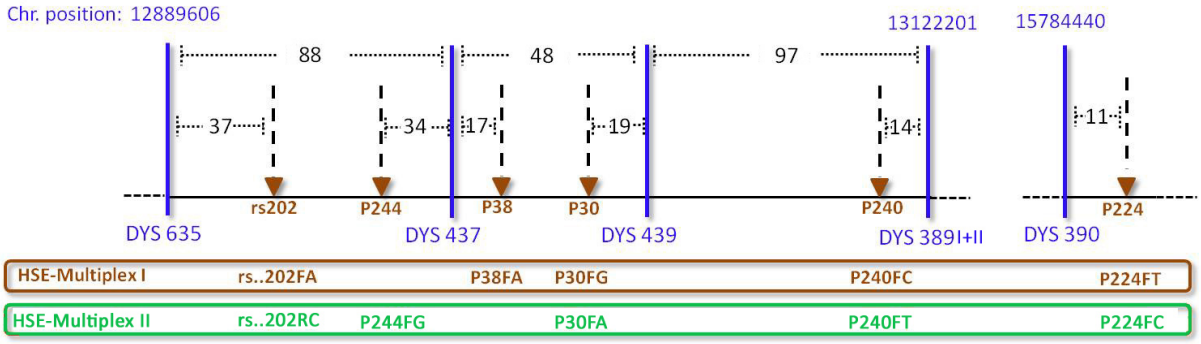
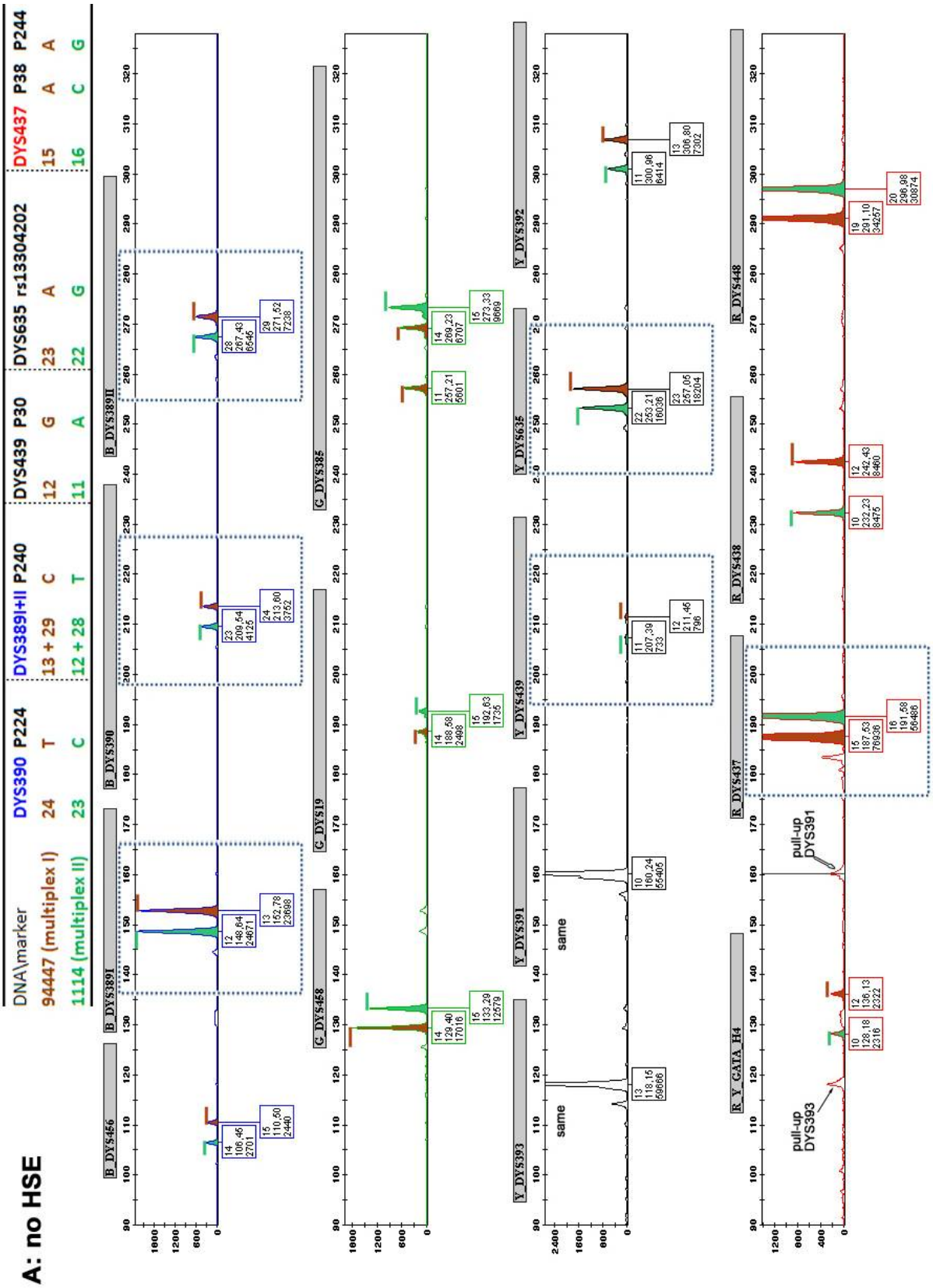


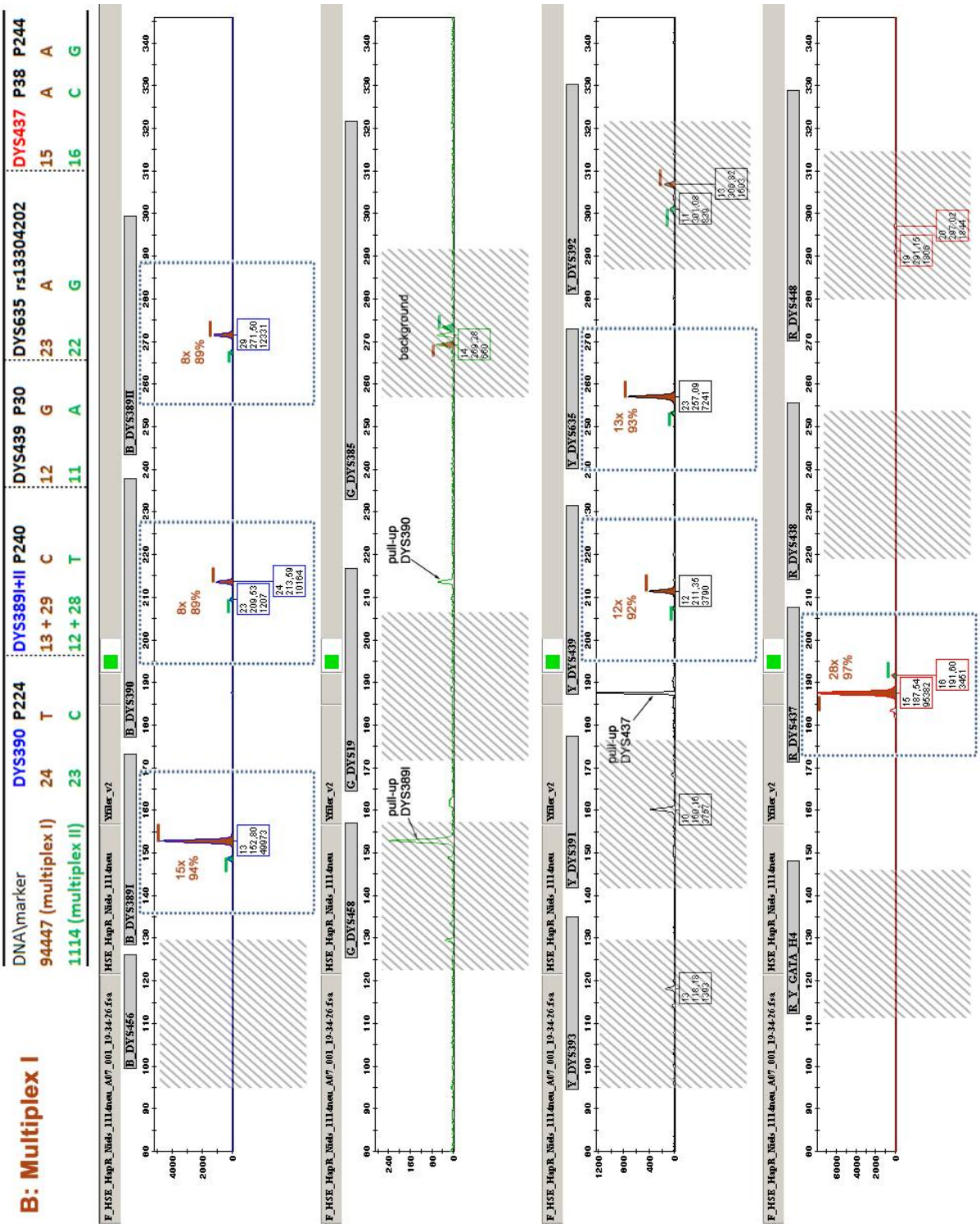
Figure 3-15: Schematic map of Y-chromosomal extraction loci of HSE multiplex I and II.
Colors indicate Y-STR markers (blue) and SNP positions which serve as the extraction sites (brown). Numbers in black show distances in kb.

The combination of several individual probes in one multiplex reaction was chosen such that within one multiplex reaction, each SNP-specific probe corresponded to the same contributor of the mixture (1114/94447). Multiplex I can extract DNA 94447 and multiplex II the DNA 1114. Hereby, multiplexes I and II are transferable also to certain other DNA mixes because all SNPs (or extraction sites) belong to one phylogenetic branch (except for rs13304202) and therefore are known to only occur in this combination. Furthermore, single probes for multiplex HSE were chosen from best HSE results obtained in singleplex HSE (Appendix-Table 7-3) All probe mixes used in multiplex reactions

were checked for the formation of possible probe dimers and potential binding sites in proximity of other extraction loci (1 kb) with the programs Visual Omp™ and Oligo Analyzer. These analyses did neither show any particularly strong formation of probe dimers nor any strong hits close to the other extraction loci. Moderated binding was observed only between the probes P240 and rs13304202. Here, the probes P240FC and rs13304202FA in multiplex I annealed with a complete match at the 3'-end of probe rs13304202FA and with -6,49kcal/mol. In multiplex II they did not show a matched 3'-terminus but bound also with relatively high ΔG values (-6,64 kcal/mol). The simulation of unspecific binding near all used extraction sites indicated a relatively high unspecific binding for the probes P38, P244, and rs133024202 near the targeted sites (between -2 and -5 kcal/mol), whereas all other probes showed unspecific binding with a ΔG under -2 kcal/mol (data not shown).

All multiplex HSEs were carried out and analyzed according to standard conditions and show successful enrichment of one contributor in all six markers intended for separation. In the majority of the tested multiplexes, the non-separated alleles showed either a drop out or appeared only as a stutter peak (Figure 3-16, blue boxes).


A: no HSE
DNA\marker DYS390 P224 DYS389I+II P240 DYS439 P30 DYS635 rs13304202 DYS437 P38 P244
94447 (multiplex I) 24 T 13 + 29 C 12 G 23 A 15 A A
1114 (multiplex II) 23 C 12 + 28 T 11 A 22 G 16 C G
B_DYS456
B_DYS389I
B_DYS390
B_DYS389II
G_DYS458
G_DYS19
G_DYS385
Y_DYS393
Y_DYS391
Y_DYS439
Y_DYS635
Y_DYS392
R_Y_GATA_H4
R_DYS437
R_DYS438
R_DYS448
same
same
pull-up DYS393
pull-up DYS391


B: Multiplex I
DNA\marker
94447 (multiplex I)
1114 (multiplex II)
DYS390 P224
24 T
23 C
DYS389I+II P240
13 + 29 C
12 + 28 T
DYS439 P30
12 G
11 A
DYS635 rs13304202
23 A
22 G
DYS437 P38 P244
15 A A
16 C G
F_HSE_HapR_Nids_1114neu_A07_001_19-34-26.fsa
HSE_HapR_Nids_1114neu
YHlex_v2
B_DYS456
B_DYS389I
B_DYS390
B_DYS389II
G_DYS458
G_DYS19
G_DYS385
Y_DYS393
Y_DYS391
Y_DYS439
Y_DYS635
Y_DYS392
R_Y_GATA_H4
R_DYS437
R_DYS438
R_DYS448
15x
94%
8x
89%
8x
89%
pull-up
DYS389I
pull-up
DYS390
background
12x
92%
13x
93%
pull-up
DYS437
28x
97%

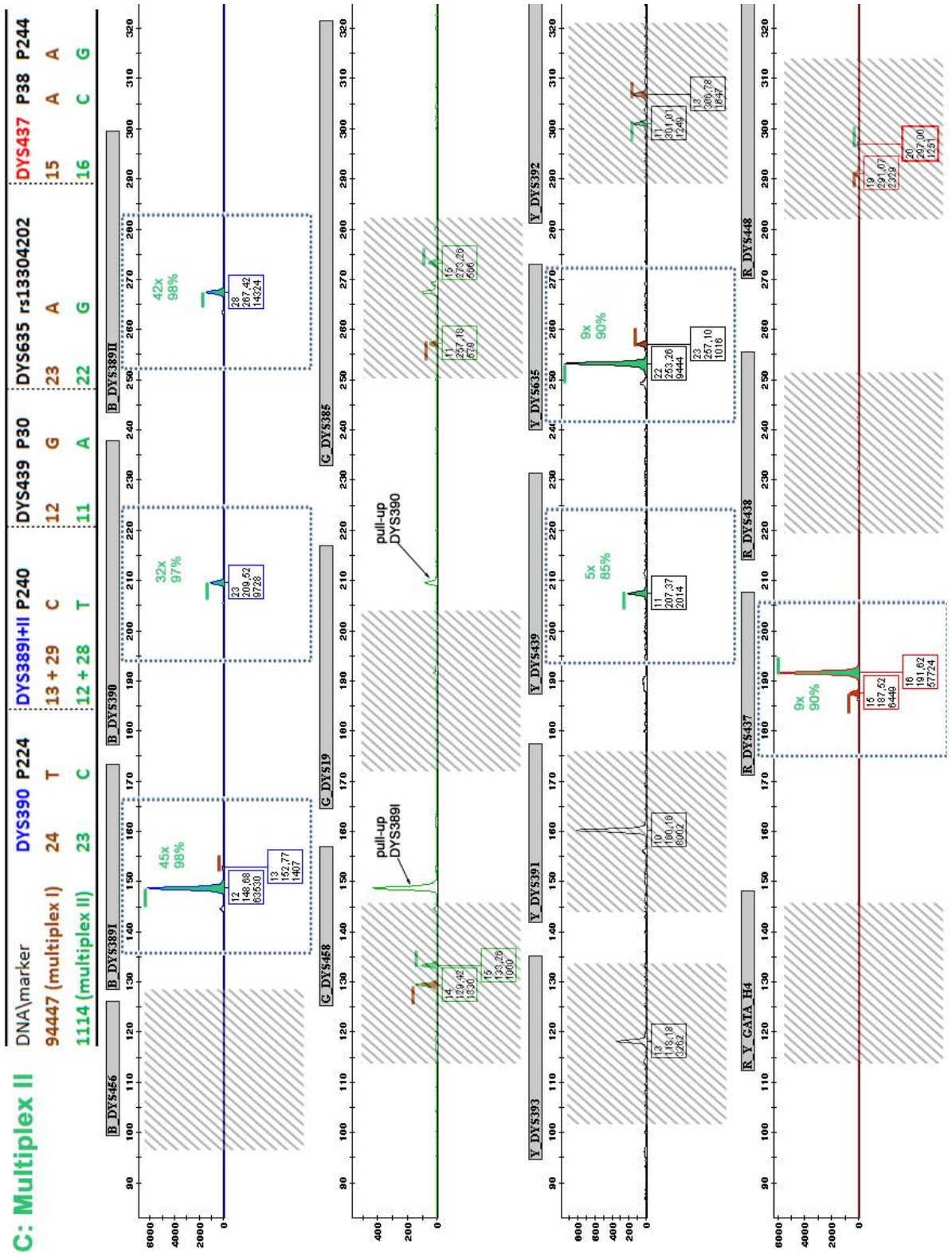


Figure 3-16: Y-filer analysis after multiplex HSE I and II.
Electropherograms for (A) the unseparated DNA mix, (B) after multiplex I to extract markers of DNA sample 94447 (brown) or (C) after multiplex II to extract fragments for DNA sample 1114 (green). Alleles of the separated DYS markers (blue dotted boxes) are colored green or brown to indicate each contributor. HSE success is given as ratio and enrichment (in %). Non-extracted markers (grey striped boxes) were detected as HSE background or failed to be detected, as expected.

3.5.2 Test of HSE multiplex design using the new prediction model at a further locus in the genome

The successful establishment of a HSE multiplex reaction for the Y-chromosome (chapter 3.5.1), should be also tested in combination with the new evaluated prediction model for probe specificity (chapter 3.4.2). In cooperation with Dr. Johannes Dapprich of Generation Biotech LLC, HSE probes were designed for the combined extraction of several human HLA loci (HLA-A, HLA-C and DRB1) as part of a project for haplotype resolution of the entire HLA locus by next generation sequencing. Target sites were selected according to next generation sequencing data of the standard DNA NA12878 in relatively conserved (i.e. less variable) intron- and untranslated regions. Due to this limitation only one or two mismatches were the only options for probe design. The DNA sample, the sequencing data as well as information about the HLA-genotype of the sample were provided by Generation Biotech. No information was given about the parental allele constellations. Probe designs for the selected target loci were based on the new prediction model (maximal Δcon) and then used directly together with the standard DNA NA12878 in multiplex HSEs. HSE samples were analyzed by Sanger sequencing of HLA-A (exon 2 – 4), HLA-C (exon 2 – 4) and DRB1 (exon 2). The combination of the extraction probes were chosen randomly since the distances between the extracted loci are larger than 100 kb. First tests of the multiplex HSE for the separation of several HLA loci showed clear enrichments for all tested loci and a complete separation was observed for one haplotype *11:01:01 of HLA-A (Table 3-7).

Therefore, the prediction of the probe specificity was very successful, considering that these results have been obtained after the first test of the HLA multiplex set up without further probe evaluation.

The sequencing results of the separated HLA alleles agreed with all pre-typed HLA alleles, except for the DRB1 locus. Here, the provided data had determined DRB1 as *01:02:01 and *03:01:01:01, whereas the sequencing results of the HSE samples only confirmed the allele *03:01:01:01 but showed *01:01:01 as the second allele.

Table 3-7: Summary of multiplex HSE at HLA locus.

communicated information from next generation sequencing data			results of sequencing after multiplex HSE					
			HLA-multiplex I			HLA-multiplex II		
HLA locus	noted for	position of HSE probe	probes	extracted alleles	HSE success [1]	probes	extracted alleles	HSE success [1]
HLA-A	no information	Intron7	A_RA+1	*01:01:01	4x	A_RG+1	*11:01:01	complete [2]
HLA-C	*01:02:01, *07:01:01	3'UTR	C_FAT+1	*01:02:01	6x	C_RAC+1	*07:01:01	6x
DRB1	*01:02:01, *03:01:01:01	Intron2	DR_RAA	*01:01:01	3x	DR_RCC+1	*03:01:01:01	3x

Default information was obtained from next generation sequencing results of Generation Biotech for standard DNA NA12878. (1) HSE success was estimated based on the ratio between peak heights. (2) Second allele was not longer detectable.

3.5.3 Required amount of DNA for successful HSE

The newly developed multiplex HSE and its separation success were also tested in dependence of different DNA concentrations. Therefore, as already described in chapter 3.2.2.1, starting DNA mixtures were created by mixing equal amounts of the male DNAs “1114” and “94447”, and then measured with the spectrophotometer. Next, mixtures were stepwise diluted. Each dilution was measured with the RotorGene Q instrument. Afterwards, multiplex HSE was carried out with the probe mixes for multiplex I and II (chapter 3.5.1) and the separation effect was calculated as the enrichment of one contributor after Y-filer analysis. Table 3-8 summarizes all results obtained: The average HSE success for multiplex HSEs with 300 ng and 150 ng DNA input was between 85% and 92%. Hereby, the maximum values for both multiplexes (300 ng and 150 ng DNA input) show always complete enrichment (>89%) and worst enrichment showed still separation effect over 70%.

Multiplex HSEs with only 50 ng of DNA input decrease in the separation success down to 67/69% for the markers DYS635 and DYS437, which correspond to the targeted sites P244 and rs13304202. The smallest enrichment obtained for these markers showed that in some extractions using 50 ng DNA input, no separation occurred. Also for the marker DYS439, some multiplex HSE reactions showed no separation. HSE with 50 ng DNA input showed always successful separation for the markers DYS389I+II and DYS390. With a further decrease of the DNA input in multiplex HSE from 20 ng to 5 ng the separation success showed higher fluctuations. Here, no separation occurred for all analyzed markers, whereas the mean enrichment remained as significant, in particular the target sites P224 and P240 (DYS390 and DYS389I+II). With a DNA input of 5 ng only the markers DYS390 and DYS389I+II showed a significant mean enrichment. Absolutely no separation for any marker was observed if only 1 ng DNA input was used.

In summary, 10 to 5 ng DNA input, which would mean 5 to 2.5 ng of each contributor, can be regarded as the absolute minimum at which successful separation can occur. Hereby the markers DYS437 and DYS635 showed a stronger decrease than the other markers. The different success rate of the markers could be influenced by several factors such as the different distance between extraction site and detected marker, the balance of the single STR marker in the multiplex Y-filer PCR or by the probe specificity itself. For example the markers DYS437 and DYS635 are more distant (> 20 kb) to their extraction loci rs13304202 and P244 than the markers DYS389 and DYS390 to their extraction sites, which could cause the stronger decrease in the separation success. In addition the extraction probes P244 and P38, which are used for the separation of the marker DYS437, indicate also a reduced probe specificity by a higher number of unspecific matches in the near of the extraction site (Table 3-3).

Analysis of Y-filer multiplex PCR with different DNA input show first drop outs and therefore also stochastic effects with DNA input of 125 pg and less (chapter 2.14.4). Hereby, the obtained peak areas drop down to 1000 and less. HSE analysis showed similar background detection for reaction with 5 ng or less DNA input. Although the detection of the HSE background cannot be directly compared with the amount of the extracted DNA it indicates the increased possibility of stochastic effects due to LCN DNA samples.

Table 3-8: Summary of HSE multiplex extractions with different amount of DNA-mixtures.

STR marker	mean enrichment in % after different amount of DNA in multiplex HSE						
	300ng	150ng	50ng	20ng	10ng	5ng	1ng
DYS389I	92 ±6	92 ±6	84 ±9	74 ±18	77 ±13	73 ±16	no
DYS390	90 ±7	91 ±6	82 ±8	74 ±15	73 ±8	62 ±12	53 ±7
DYS389II	92 ±5	91 ±8	85 ±8	71 ±18	77 ±11	63 ±15	no
DYS439	89 ±6	90 ±3	81 ±9	67 ±14	70 ±15	57 ±14	no
DYS635	85 ±8	87 ±11	67 ±9	58 ±15	56 ±14	49 ±12	no
DYS437	86 ±7	87 ± 2	69 ±15	68 ±14	70 ±12	57 ±17	no
n	23	4	9	11	12	18	8

The color scale distinguishes HSE success: HSE success decreases from dark green to dark orange. Given amounts of DNA used in HSE reactions are defined as: 1=1-2ng, 5=4,5-5ng, 10=7-10ng, 20=18-25ng, 50 =45-75ng and 150ng = 90ng – 170ng of total DNA input material. Note that given DNA amounts are calculated for DNA mixture and not for the single contributor. (no) no enrichment was observed, (n) number of experiments.

Together with the background analysis in chapter 3.2.2.1, the background from multiplex HSEs for decreasing amounts of DNA was also analyzed. The results are summarized in Figure 3-17, which

illustrates the background peak areas and the drop out rates after Y-filer multiplex PCR with different input DNA amounts with and without previous HSE extraction.

Figure 3-17 A shows the continually decrease of the obtained peak areas after Y-filer PCR (without HSE) with decreasing DNA input. As expected the Y-filer analysis from HSE samples with decreasing DNA input (Figure 3-17 B) result also in a decrease of the detected background. In addition, the concentration of HSE samples was measured by real time PCR, which amplifies an autosomal target and therefore should give some information about the extracted HSE background. These measurements showed a concentration of approximately 110 and 70 pg/µl for 150 ng and 50 ng DNA input in HSE, which would mean an input of 825 pg and 525 pg DNA mix in the Y-filer reaction (7.5 µl sample volume for one Y-filer reaction). Y-filer analysis from HSE reactions with 20 ng and 10 ng DNA were calculated to use about 375 pg (50 pg/µl) and 150 pg (20 pg/µl), whereas HSE reaction with 5 ng and less DNA input result in very low final DNA concentrations which allow only the use of 100 pg DNA and less in Y-filer analysis. Corresponding to the measured concentration of the HSE samples Y-filer analysis result in peak areas of 10000 to less than 1000, which correspond also to used DNA concentration and obtained peak areas for the Y-filer analysis without HSE.

In summary, the background signals of HSE with different amounts of DNA indicate that maximum obtained peak areas depend mainly on the amount of input DNA in HSE reaction, whereas low background signals can also occur in HSE reactions with high DNA input. Concluding, it seems that the non-target loci are extracted rather inconsistent during HSE, which can be also seen by the drop out rates. Here, the tests of the Y-filer PCR with DNA dilution series showed first drop outs with a DNA input of 70 pg and less. In comparison, Y-filer analysis of HSE samples showed inconsistent drop outs rates.

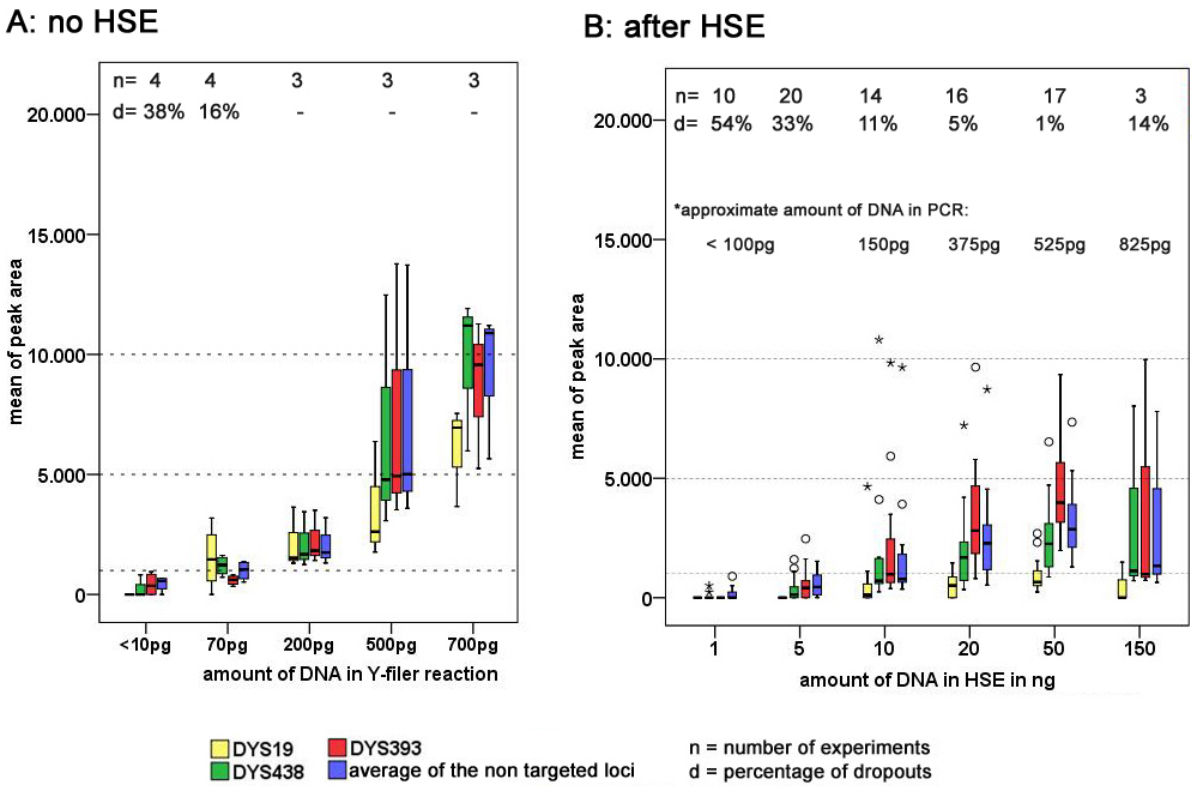


Figure 3-17: Signals of the non-targeted Y-filer loci with different amounts of DNA without and with HSE.

The diagram shows the dependence of the peak areas obtained for the amount of input DNA after (A) direct Y-filer analysis of a dilution series and after (B) Y-filer analysis of HSE reactions with different amounts of DNA input. For all reactions, the male DNA mixtures 1114/94447 and 1118/5573 were used. DNA amounts were determined by the measurement of DNA dilutions with the Rotor-Gene Q and the spectrophotometer. (*) DNA-background concentration of HSE samples were measured with the Rotor-GeneQ and theoretical DNA input in Y-filer reaction was calculated for 7.5 µl sample volume. Given amounts of DNA used in HSE reactions present groups which are defined as follows: "1"= 1 bis 2ng, "5"=4.5-5ng, "10"=7-10ng, "20"=18-25ng, "50"=40-50ng and "150"=150–170ng. HSE reactions were carried out with probes for the target locus M173 or probes for multiplex I and II. Note that given DNA amounts are calculated for DNA mixture and not for the single contributor. Dotted lines indicate a peak area of 10000, 5000 and 1000.

4 DISCUSSION

In this work the HSE technique was applied for the first time to physically separate human male DNA mixtures for further forensic investigation. This means that the already existing HSE approach, known for the separation of HLA haplotypes, needed to be adapted for the specific extraction of large, human Y-chromosomal fragments. The first step for the establishment of a new HSE approach was the search for suitable single- or polynucleotide polymorphisms.

4.1 Database research after variation for HSE probe design

At the beginning of this study, a search for Y-chromosomal variations as extraction loci near Y-STR markers was carried out. In particular the search was focused on polynucleotide polymorphisms because of the insufficient results of first HSE application on the separation of Y-chromosomal haplotypes (chapter 3.2.1), which used only mononucleotide polymorphism. Furthermore it was searched for SNPs near the Y-filer STR markers DYS391, DYS390, and the group of DYS635 to DYS389I+II. The starting point for the SNP search was the NCBI database, because this database covers most of the submitted SNPs, supplies additional information, such as submitter details, mapping position or allele frequencies and contains convenient search tools. Nevertheless, other databases were also used, including the International Hap Map Project, 1000 Genome Project, YCC, YHRD, and the Family Tree database. The involvement of further databases helped to uncover and exclude most of the SNPs with ambiguous mapping. The Mammalian Genotyping database was especially used to search for polynucleotide SNPs. After this first selective step, the only SNPs that remained had unambiguous mapping positions and matched the criteria of interest.

Next, paralogous sequence variants (PSVs) were eliminated. The first indication for the presence of many PSVs is an elevated SNP density, such as an average SNP density of 0.69 to 1.33 per kb in duplicated regions [148]. In this search, SNP densities of approximately one SNP per 100 bases were found for several regions. The MID504 variation, located on the X-transposed region (XTR) of the Y-chromosome, is one such region. It has 98.78% sequence identity with the X-chromosome [173]. As shown in Figure 3-1, the MID504 region represents, according to the NCBI database, a high density of SNP entries, which are most likely PSVs. The best method for identifying a PSV is to obtain a BLAST analysis of each SNP spanning region. The region of MID504 showed next to the match on the Y-chromosome, a second hit on the X-chromosome. Furthermore, the sequence difference between the chromosomes at the MID504 locus (X chromosome: ACA insertion; Y-chromosome: ACA deletion) was identical to the base substitutions of the database entry (-/ACA). This was also observed for the

eight additional entries on the alignment fragment (Figure 3-1). Therefore, the BLAST results strongly indicated that all the supposed SNP entries attributed to the MID504 region really are PSVs.

This was rather surprising, because the MID504 data were based on a comprehensive population study of 430 chromosomes and one would have expected this fact to become clear during such a large study. The sequencing results of this region did not reveal any variations in MID504 or in the four additional loci (rs2757249, rs2534929, rs2534930 and rs2534931) included in the test with MID504. Consequently, those database entries were proven to be invalid. Moreover, the validity of the untested SNP entries in the MID504 region remains highly doubtful (Figure 3-1). This suggests that the submitted allele frequencies for MID504 (Figure 2-1) must be considered as erroneous data.

Another important indication for PSVs is the method with which the variations have been discovered. Computer-based discovery strategies tend to generate numerous PSVs, because they often accidentally categorize nucleotide mismatches in homologous regions as SNPs, but not as PSVs. However, even after excluding PSVs, it still remains possible that a SNP entry might be caused by a sequencing error or a private SNP. The only way to conclusively recognize these errors is to check all the available and validated information. The NCBI websites summarize the submitted information and create a validation status for each SNP. As of June 2011, the NCBI also released new guidelines for a better recognition of private SNPs and PVSs. For example, at the time of this study the SNP rs2757249 was listed as “valid” in the NCBI database (Build 36, 2009), but was shown to be a false entry. At present rs2757249 is labeled as “suspect” in NCBI database. However, for reasons unknown, the SNPs rs2534929, rs2534930 and rs2534931, which are very likely to be PSVs, are not labeled as “suspect” in the dbSNP database at the moment. Although the new attributes do improve the database information, the evidentiary value of each “valid” SNP should be regarded highly critically [174].

In general, experimentally-based discovery strategies provide higher quality than computer-based analyses, and comprehensive population studies and heterozygosity data are the best certification for a true SNP. Nevertheless, this high quality data can also contain errors, as it could be demonstrated in the case of MID504.

This error might arise for two reasons. First, the typing may very likely have been carried out with no Y-specific PCR primers; consequently, the PCR products might be obtained from the X-chromosome and from the XTR on the Y-chromosome. Second, the PCR may have been carried out on a DNA pool that combined male and female DNA; consequently, it would be impossible to distinguish between female and male DNA results.

In this study, the variance MID504 was identified as an Indel, which distinguishes the X- and Y-chromosomes from each other; however, there was no variation on either the Y or the X

chromosome. In other words, the ACA insertion only existed on the X-chromosome and the deletion was only found on the Y-chromosome. The transition of the XTR region, in which MID504 is located, occurred approximately 4.7 Myr ago through a duplication of material from the X to the Y chromosome, after the divergence of the human and chimpanzee linage [90, 96, 173]. BLAST analysis of the Y-chromosome sequence of the MID504 region with chimpanzee and *rhesus macaque* genome showed only a hit on their X-chromosome. In contrary to the human X-chromosome, no ACA insertion could be seen for the ape X-chromosome. Therefore, it seems that after the transfer of X-chromosomal region an insertion of a second ACA repeat occurred on the human X-chromosome.

Interestingly, the SNP rs5940712, which is located exactly in the center of the ACA insertion on the X-chromosome, represented an additional sequence change inside the ACA motive (which must have formed later). In this study, the SNP rs5940712 could be identified as a true variation which is in agreement with the database information. Here the BLAST result did not indicate the rs5940712 as a PSV, and the NCBI database showed several validations and a comprehensive population study.

The most frequent problem is that SNPs are annotated with an exact mapping position, but have unconfirmed status with little or no additional information. In this study further investigations were carried out for three of such unconfirmed variations (rs35753737, rs34485380, and rs33963329; Table 3-1). None of these three entries showed any indication of a PSV, but they were reported in the NCBI to be discovered in very small sample sizes. The variations rs35753737 and rs34485380 were reported to be found by sequencing Celera Donors which included only two male individuals with unclear ethnicities. The entry rs33963329 was reportedly found by re-sequencing of only two chromosomes. Because the typing results from at least 64 individuals with different ethnicities did not reveal any variations, it was concluded that the two database entries were not true SNPs; instead, they might have arisen due to sequencing errors or private variations. It is of course possible that a larger study with more individuals might reveal variations.

The results for the variations rs35753737, rs34485380, and rs33963329 were submitted to the NCBI database to provide more information. The variation rs33963329 is most likely a mistake caused by an error in sequencing or alignment. Although it was discovered by re-sequencing, rather than by a computer-based approach, the error could be due to its location adjacent to the repeat element DYS718. An alignment of sequences that differed in the number of TTA repeats in the DYS718 locus could have caused a false SNP identification. Support for this mechanism can be found in the BLAST result for rs33963329. Here, the alignment of the reference sequence with the Y-chromosome alternate assembly sequence showed one repeat variance (Figure 4-1). In addition, the DYS718 locus was described as heterozygous for Chinese Han Populations [175], and was determined to be

variable in this study with the observations of: five person with $[TTA]_{15}$, one person with $[TTA]_{14}$, and two person with $[TTA]_{15}$ repeats.

```
Homo sapiens chromosome Y genomic contig, alternate assembly

Query  301     AATCCATATGTAAGACACAATGCAAAAAGAAAGAAGAATAAACACTATTATCTGTGTTAG  360
               ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  833933  AATCCATATGTAAGACACAATGCAAAAAGAAAGAAGAATAAACACTATTATCTGTGTTAG  833992

Query  361     TATTATCTTAGGAGAAAATTCAATGCAGTTACCTATTTCTNTATTATTATTATTATTATT  420
               |||||||||||||||||||||||||||||||||||||| ||||||||||||||||||||
Sbjct  833993  TATTATCTTAGGAGAAAATTCAATGCAGTTACCTATTTCT-TATTATTATTATTATTATT  834051

Query  421     ATTATTATTATTATTATTATTATTATACTTTACGTTTTAGGGTACATATGCACAATATGC  480
               ||||||||||||||||||||    ||||||||||||||||||||||||||||||||||||
Sbjct  834052  ATTATTATTATTATTATTATTA---TACTTTACGTTTTAGGGTACATATGCACAATATGC  834108
```

Figure 4-1: BLAST result of the locus that harbors SNP rs33963329 on Y-chromosome.
The BLAST result for the flanking sequences of rs33963329 (Query, N = -/TTA) showed only one hit on chromosome Y, close to a simple repeat element, DYS718 [175], with the same 3-base pair motif (TTA, highlighted in red and green). Furthermore, the Y chromosome alternate assembly sequence (Sbjct) had one less TTA motif (blue) than the reference sequence (Query).

After the unsuccessful search of polynucleotide variations near Y-filer STR markers, the search was extended also to mono- or single nucleotide polymorphisms, which are more frequent. Here, in total 14 SNPs (excluding the SNP rs595012, Table 3-1) were investigated which were already well characterized in the database. Additionally, two SNPs were added based on former laboratory data. Two additional single nucleotide variations were found as private SNPs near the marker DYS437, S4:2701 and S6:4204. From these 16 analyzed SNPs, 13 SNPs were used for probe design and therefore 13 extraction loci were available for further investigation of HSE. Together, these loci showed distances between 150 bp and 243 kb to nearest Y-filer STR markers. However, it should be noted that these SNPs obviously present only a selection of the available database entries.

4.2 Separation of male DNA mixtures by the HSE technique

First attempts to transfer the HSE technique to the separation of mixed Y-chromosomal marker with subsequent STR-typing showed only little success. Results showed indeed an enrichment of each of the appropriate contributor, but only to a little extent (chapter 3.2.1). The enrichment obtained was only between 1.2x and 1.7x, which remained doubtful to be significant, especially for the internally set up markers DYS495, DYS635 and DY504. Therefore, for further analysis, it was focused only on the Y-filer markers because of their more balanced and sensitive PCR primers and the advantage of multiplex amplification of several markers. For a better and more precise assessment of the significance of the separation effect, Y-filer analysis with prepared DNA mixtures were carried out to evaluate kit specific imbalances (chapter 2.14.4). This analysis showed that peak fluctuations, which

occur up to 61% enrichment could be caused by Y-filer PCR or DNA mixture imbalances alone, whereas enrichment of over 70% can be considered as a reliable separation. In summary, initial test of the HSE technique for the separation of male DNA mixtures showed no remarkable success. Additionally, the investigation of the enrichment of the marker DYS390 with the Y-filer kit showed also background signals for non-targeted loci. Reasons for the few separation success could be a low probe-specificity. Typical HSE probes for the separation of HLA haplotypes take advantage of two, three and often more polymorphisms due to the high variability (available) at this locus. Another reason for the poor separation success probably was the unspecific binding of genomic DNA to the beads, which caused the high levels of background signals in the sensitive Y-filer assay.

4.2.1 Optimization of the hybridization buffer

One key component of the HSE reaction is the hybridization buffer. Therefore, studies were undertaken to optimize the main buffer components: *Taq* polymerase, dNTPs and biotin-dUTP by testing different concentrations (Figure 3-5). These studies revealed that changes in dNTPs or biotin-dUTP concentration did not improve HSE reactions. Interestingly, there was a broad range of biotin-dUTP concentrations for which specific DNA extraction by HSE occurred.

Comparison between the available binding capacity of the beads and the actual amount of biotin used during HSE showed that only 2% of the total beads binding capacity is utilized (detailed information cannot be provided due to confidentiality restrictions). The number of genomic copies for one contributor in one HSE reaction can be calculated as follows:

$$number\ \text{of molecules}_{\text{Target-HSE}} = \frac{m_{\text{Target}}}{MW_{\text{Target}}} \times 6.02\text{x}10^{11} \qquad (21)$$

Whereas the MW_{Target} for human template DNA can be calculated after equation 10 (chapter 2.13), m_{Target} is the amount of input DNA (approx. 0.15 µg DNA for one contributor) and 1 pmol = $6.02\text{x}10^{11}$ molecules (Avogadro constant). Using equation (21), approximately 43000 haploid genome copies can be calculated for one contributor. The incorporation rate of biotin-dUTP molecules during the extension of the target extraction locus as well as the length of the extension product is not known. The degree of biotinylation was analyzed by Theissen *et al.* 1989 and Fenn *et al.* 1990 [176, 177], who studied the incorporation of Bio-4-UTP or Bio-11-UTP into RNA by invitro transcription with T7 RNA polymerase. Both groups found that the number of incorporated biotin molecules depend on the Bio-UTP to total UTP ratio during transcription. However, both groups reported different incorporation efficiencies. The group of Theissen measured that Bio-UTPs are incorporated three to four time less efficiently than unmodified UTPs, whereas Fenn and Herman showed a nearly 50%

biotinylation for reactions with equimolar amount of UTP and Bio-4-UTP. If assuming in a simplified model that all four different bases occur in the target sequence at the same frequency based on Theissen and Fenn an incorporation of biotin residues can be estimated every 8th till every 32th nucleotide during HSE with equimolar amount of dTTPs and Bio-dUTPs. The manufacturer's specifications of *Taq* polymerase enzyme allow a maximum amplicon size of 5 kb. Based on this length approximately 625 biotin residues would be incorporated in one Y-chromosome target during HSE reaction when using the highest estimated efficiency for biotin incorporation and polymerase extension. It has to be noted that during extension reaction the maximum product size of a newly synthesized strand might be longer. In total, only about 2.5 million biotin-dUTPs molecules would be needed to completely label all input Y-chromosome target contributor DNA with biotin (625x43000), which would correspond to less than 0.1% of the input biotin used in the extension reaction (detailed information cannot be provided due to confidentiality restrictions). Although these are quite speculative calculations, they show that HSE extension reactions are carried out with a large excess of biotin. This could explain why no significant change in separation success for lower biotin-dUTP concentrations could be observed, because sufficient biotin-dUTPs would still be available for the reaction. On the other hand the results showed that the separation success decreases significantly with 5x and 10x more concentrated biotin which indicates that the biotin concentration arrived at critical values at which specific extension reaction can occur.

The polymerase turned out to be the most important component in the HSE chemistry. Reducing the polymerase concentration to 0.5x or 0.25x could, on average, double the separation effect (Figure 3-5 D), which also confirms the results that were obtained by diluting the entire HB (Figure 3-5 A). These results are in agreement with the observation of Sommer *et al.* 1992 [178] and Bottema *et al.* 1993 [40], who showed an increased specificity of allele-specific amplification by lowering magnesium or *Taq* polymerase concentration. Furthermore, an improved fidelity of the *Taq* polymerase has been demonstrated for lower concentration of dNTPs, $MgCl_2$ or lower pH, whereas polymerase responded strongly to changes in $MgCl_2$ concentration and pH [179]. Despite the new optimized composition of HB, however complete separation of the contributor alleles was not achieved. Therefore, further investigations of HSE probe specificity were carried out.

4.2.2 Probe specificity

The second part of the HSE optimization was focused on probe specificity, since the changes of the hybridization buffer did not result in complete separation of one contributor. Problematic on the Y-

chromosome is the absence of polynucleotide polymorphisms near Y-filer markers, which would increase probe specificity. Although allele-specific PCR (ASP) and allele-specific extension reactions are difficult to compare, ASP was used as a first test of HSE-probe specificity by using the allele specific probe as a primer in PCR. As expected the results of the ASPs, showed that the polymerase failed allelic discrimination depending on the annealing temperature (data not shown). Therefore, HSE were also tested with different hybridization temperatures. Results showed indeed an increase of HSE success for higher hybridization temperatures but to a very small extent, which was even lower than the improvement HSE success obtained with reduced polymerase concentration (data not shown). The reason for the negligible improvement of the separation success could be: First, the higher hybridization temperature (T_{Hyb}) could inhibit the re-hybridization of large genome fragments and second, higher T_{Hyb} would also increase polymerase activity which in turn could worsen the separation effect as discussed before (chapter 4.2.1).

4.2.2.1 Probe design based on the new prediction model

HSE probe design had been already established for the separation of HLA haplotypes. However, this application design contains some substantial differences. First, the target loci and extracted sequences are located close to each other and second, the high variability of the HLA locus allows the design of probes with several mismatches. On the Y-chromosome the extracted loci and the analyzed markers show several kb distances, and probe design is in general restricted to only one mismatch, which reduces mismatch discrimination. Probe design therefore is more complicated and should be studied in more detail. Because primary probe design is mainly limited to the orientation and length of the probe, first different probe length in both orientations were tested.

The results of HSE with different probe lengths clearly demonstrate that length differences of only few base pairs can determine HSE efficiency. For example the probe P224 FC showed 63% separation efficiency for a length of 23 nucleotides. When four nucleotides were removed from the 5'-end of the probe, the separation efficiency increased to 98% (Figure 3-6).

The dramatically improved separation of the two male samples with the 19-nucleotide long HSE probe was unexpected, as shorter probes generally display lower sequence specificity for their targets as well as higher dissociation rates. This same effect was observed with a number of probes (Figure 3-8 and Appendix-Table 7-3). Although the 3'-termini were consistent in all evaluated probe sets, the improved separation with altered probe length was not likely caused by increased mismatch recognition by the polymerase. Other probe-associated factors, such as G/C content, melting temperature, and Gibb's free energy (ΔG), also were non-determining factors that could be

correlated to HSE efficiency. Comparing the simulated probe concentration between match and mismatch hybridization ($\Delta con_{M\text{-}MM}$) for each probe set with HSE separation success, a strong correlation between maximum $\Delta con_{M\text{-}MM}$ and HSE efficiency was observed. In more than 80% of the comprehensively tested probe sets, the length of the best HSE probes consistently matched the predicted best probe length with no or only one base difference. The data show that Δcon between match and mismatch represents a potential predictor of probe specificity. It also indicates that the specific elongation of the 3'-end does not only depend on the mismatch discrimination by the polymerase, but – importantly - also on the ratio of available targets for match and mismatch.

The relevance of the difference in concentration between the matched and mismatched probes in the HSE experiments is independent on the target concentration, since no amplification of the target occurs in these reactions and the probe concentrations are in the order of 10^9 more concentrated than the intended genomic DNA target. The data suggest that these concentration differences between probes and targets promote a highly competitive environment between match and mismatch probes in which small changes between concentrations of matched and mismatched probe apparently affect polymerase activity, as $\Delta con_{M\text{-}MM}$ is the only parameter demonstrated to predict HSE efficiency. Also in the work of Fish *et al.* 2007 [180], the target concentration is supposed to be a critical experimental parameter in achieving optimal discrimination of SNPs in DNA microarrays. They suggest that a lower abundance of available target leads to a competitive dynamic between perfect matched and mismatched SNP probes.

The curve progression of Δcon and HSE success did not always completely match. For example, for probe set P224FC, separation by HSE can still occur with very small Δcon values downstream of the maximum Δcon, and the HSE curve of P30FA seems to be shifted to a probe shortened by one nucleotide. These minor changes could be caused by additional factors, such as short distance from the marker, favorable mismatch for polymerase discrimination or a more specific sequence even for shorter probe lengths.

Strong disagreement between the correlation of maximum Δcon and HSE success was observed mainly with P38. The simulated maximum Δcon was predicted to be much smaller than the length obtained for most specific probes after HSE experiments. A possible cause of this observation may be the high observed hit frequencies after the alignment of P38 in the 30 kb distance, which in turn might be caused by its high G/C contents. Guanosines are promiscuous and probes with low complexity that contain runs of three or more guanosines in a row tend to mishybridize to either its intended or unintended target [48]. Furthermore, the increased non-specific binding of P38 could cause rapid loss of the probe pool and influence final match and mismatch hybridization which

cannot be considered in the simulation. The simulation of P38 probe hybridization with lower starting concentration showed a switch of maximum Δcon to longer probe length and therefore, could explain the observed results. Importantly, Δcon only provides information about the proportion of match and mismatch. This factor influences polymerase discrimination in HSE strongly but mismatch discrimination further depends on polymerase specificity. Therefore, it can be concluded that the unexpected separation observed in HSE for probe rs13304202FA probably resulted from favorable mismatch recognition by the polymerase.

This study showed that the ability of the polymerase to discriminate between matched and mismatched probes was the highest (in 88% of the cases) when the ratio of match to mismatch targets was at a maximum. The obtained data clearly prove that probes for allele specific extension reactions work only within a narrow and variable range of probe lengths. Figure 4-2 presents the occurrence of Δcon for different probe length at the example of the probe set P224 FC. Hereby the duplex concentrations, which increase with increasing with probe length, are blotted against Δcon. Therefore, Δcon is illustrated in double form: first, by the Δcon values itself (Y-axis) and second, by the switch of the Δcon values depending on the occurring duplex concentration for match or mismatch. For example, the blue square and triangle present the same maximum Δcon value, obtained for probe P224 (19nt long). Hereby, maximal Δcon for match duplex is shifted to higher concentration of hybridizated probe, because of the more stable match. With longer probes Δcon decreases but shows higher concentration of hybridized probe for match and mismatch (red).

For shorter becoming probes Δcon also decreases slowly but exhibits decreasing duplex-mismatch concentration and therefore less unspecific probe binding (green). Hereby, the minimum probe length seems to be determined by several factors:

a) decreasing probe length provokes less binding stability and reduced concentration in probe-target duplexes, which affects also primer extension efficiency

b) with shorter probe length, the probe becomes less sequence specific,

c) Δcon between match and mismatch becomes insignificant.

One further limiting factor for minimal probe length of course is the minimal primer length required for the binding of the polymerase. Primer extension analysis showed that both the KF and the *Taq* polymerase require a primer of at least seven or eight nucleotides respectively and that minimal primer lengths corresponds to the distance between the catalytic center of the “palm” and the “tip” of the “thumb” domain [181].

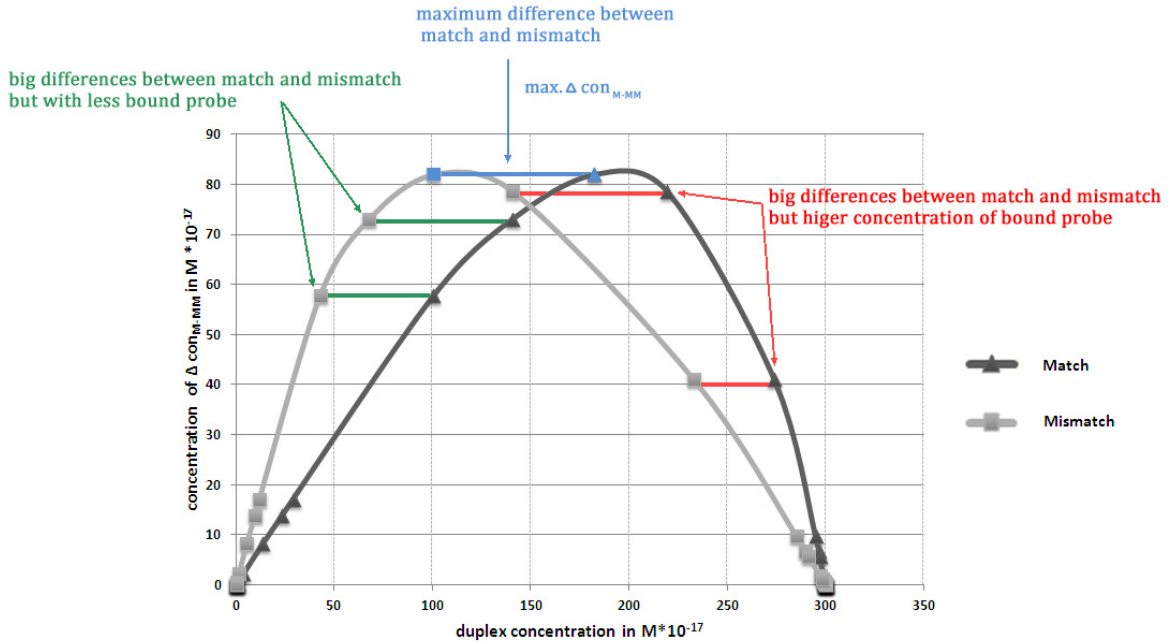


Figure 4-2: Plot of Δcon against duplex concentration of match and mismatch for the probe set P224 FC. Diagram shows Δcon for increasing duplex concentration for match (dark grey triangles) and mismatch probes (light grey squares) with different probe length.

It has to be considered that HSE includes the formation of considerable homo- and hetero-duplexes on the genomic level. Therefore, hybridization temperature and time were set to 58°C for 20 min to allow rejoining of the complement strands, but this can also limit the available annealing temperature range for probe-specific binding. Furthermore, in HSE reactions the entire human genome serves as a template for probe binding. Although a complete match of an average HSE probe (length 15 -18 nt) is expected to occur only once in the entire human genome, the real number of matches can be increased dramatically through duplications, transpositions, and conserved sequence motifs in the genome. Therefore, binding sites for partially annealed probes typically occur in the genome for an unmanageable number and due to the low annealing temperature of 58°C the amount of non-specific binding also strongly increases, whereas specific binding decreases with shorter probe length. On the other hand, the maximum probe length, as determined by the prediction model, only seems to depend on the concentration difference between match and mismatch duplexes. Importantly, when the probe length is longer, mismatch hybridization increases and thus still lead to a high undesired extension of mismatched targets (Figure 4-2).

In addition to the correlation of maximum Δcon and HSE success, it has been observed that the simulated Δcon partly depends on the G/C content of the probe and that increasing G/C content causes a shift of Δcon to shorter probe lengths. The same dependency of G/C content and probe

length was observed for the best specific probes tested in HSE and confirms the prediction model for best probe length after match and mismatch hybridization with Visual Omp™.

4.2.3 Separation of Y-chromosomal haplotype by multiplex HSE

After the adaption and optimization of the HSE technique to the extraction of Y-chromosomal fragments of male DNA mixtures, a multiplex HSE reaction for a simultaneous separation of several STR markers of one contributor was to be established. Therefore, the best evaluated probes from singleplex HSE (chapter 3.4.2) were combined in one HSE reaction for five extraction sites. The results show that with the combination of several probes all target loci could be separated within one HSE reaction. Multiplex HSE results in the separation of a small haplotype profile with the simultaneous separation of six Y-STR markers. Various tested combination of probes in HSE did not show significant differences and indicated that the combination of several probes in a multiplex extraction reaction can be carried out without difficulties.

The obtained HSE efficiency showed mean values for one multiplex HSE between 78% and 91% of enrichment. Hereby, some obtained minimum and maximum values of the single probes lie between 70% and 98%. The minimal values are caused by loci P244 and rs13304202 which show higher fluctuation of their HSE success. The more unstable HSE results could be caused by in general unfavorable mismatch discrimination, higher distance between extraction loci and STR marker and additional unspecific probe binding. HSE separation with less robust probes might be improved by using modified oligonucleotides, such as locked nucleic acids (LNA) or a 4'-C modification like the methooxymethylene group which has been reported to enhance allele discrimination [182-185]. Multiplex reactions were also compared with the corresponding singleplex HSE since some markers targeted in multiplex HSE possess two extraction sites in close proximity. However the comparison of the mean values between single- and multiplex did not show significant differences. One reason for this might be the well-established extraction success in the singleplex HSE, which allows only little room for further improvements. A test of multiplex reactions with HSE probes, which show a poor separation success in singleplex HSE, might show a more clear improvement of the HSE success in multiplex reaction.

On the other hand, multiplex HSEs, set up with the same probe-mix, also varied in their separation success, which possibly could be caused by different storage time of certain HSE components, such as probes or DNA mix. This is discussed also later in chapter 4.2.4. Therefore, the comparison of not synchronically set up HSE reactions seems to be hampered and should be regarded only deliberately. Synchronically set up HSE reaction with same HSE master mix, extraction run, and Y-filer analysis as it

was carried out for the studies of buffer compositions and probe sets, can be regarded as best comparable unite without additional error.

The obtained enrichment reflects ratios between approximately 2 and 50 times higher detection of one contributor. How much of these calculated ratios represent the real ratio between targeted alleles and unspecific extracted alleles depends on an accurate and balanced amplification of the different target during Y-filer multiplex PCR. The Y-filer kit works only in a narrow range of DNA concentrations and gives results for all analyzed markers (complete profile) for minimal DNA input of approximately 125 pg DNA per reaction. The optimal amount of DNA input is about 1 ng. PCR reactions with less than 60 pg DNA input fail to amplify markers and result in partial or non profile, whereas PCR reactions with more DNA input result in unspecific amplification and strong background.

This narrow range of the DNA amount in the Y-filer reaction therefore sets the limit of resolution of mix ratios to about 1:20 depending on the DNA amounts of the major and minor allele [106, 167]. DNA mixtures that contain one component outside of this range could cause an unbalanced amplification during Y-filer PCR and therefore result in erroneous ratios. Therefore it is likely that Y-filer analysis of HSE samples can determine the real ratios of the two contributors with a relative high accuracy for ratios up to the Y-filer detection limit of 1:20.

One additional difficulty for the exact determination of the ratio between the extracted alleles and the unspecific background is that in approximately 50% of the extraction reaction the non-targeted alleles of the second contributor happen to fall into the N-4 position. In this situation the separation success would incorrectly appear as decreased during Y-filer PCR because of the generally higher observed peak areas for alleles in the N-4 stutter the position [66]. However this inaccuracy could not be avoided since many alleles of two contributor show a differences of only one repeat unit. A more clear description about the real concentration of the separated target and HSE background could give for example a real time approach with TaqMan probes for different extracted and non-targeted loci.

Regardless of this theoretical discussion of the real rate of targeted and non-targeted alleles in HSE samples, the tested multiplex HSE clearly results in a clear separation and identification of one contributor haplotype.

4.2.4 Investigation of background signals after HSE

Samples after HSE reaction show next to the enriched locus also often strong signals of other Y-chromosomal and autosomal loci. Supposing that high unspecific extraction could also influences HSE efficiency, the observed background was investigated based on following aspects:

- Background generated by unspecific binding of genomic DNA to the magnetic streptavidin coated beads.
- Background generated by unspecific labeling of genomic DNA with biotin-dUTP.

4.2.4.1 Background signals by unspecific binding to beads

Because undesired unspecific extraction could affect HSE success, the observed background was investigated according to additional wash steps, bead-blocking and different amounts of DNA input. The aim of the blocking experiments was to cover unspecific binding sites on the beads so that only biotinylated fragments can be extracted. The HaploPrep Kits for HSE uses Sera-Mac magnetic beads which were pretreated with a blocking solution based on proteins but without any DNA component. Bead-blocking with female DNA showed that the beads are indeed able to bind unspecific non-biotinylated DNA. However, this unspecifically bound DNA could be removed efficiently by subsequent washing of the beads. The use of blocked beads or additional washing steps in standard HSE reaction showed neither a decrease of background nor an improved separation effect. Therefore, it has to be concluded that the provided streptavidin beads as well as the standard HSE wash protocol are well adjusted.

4.2.4.2 HSE Background in HSE by unspecific labeling with biotin-dUTPs

Because the observed high background of HSE samples could not be related to unspecific DNA binding, it can be considered that unspecific biotinylation occurred during HSE reaction. Experiments with or without decreased *Taq* Polymerase and biotin-dUTP concentrations indicated the presence of unspecific labeling of the genomic DNA because background signals decreased with less concentrated polymerase or biotin. HSE without any polymerase or biotin-dUTP showed almost a complete loss of signal background. Therefore, it could be conclude that the observed background after HSE results mainly from unspecific labeling of the genome with biotin-dUTP by the polymerase rather than by unspecific binding of some DNA to the magnetic beads.

As already described in chapter 4.2.1, biotin-dUTP is used in HSE in considerable excess and only a tiny part is actually needed for the labeling of the extraction loci. Therefore, the high amount of remaining biotin as well as the high binding capacity of the beads may allow significant unspecific binding of labeled but non-targeted DNA fragments. When considering the excess of biotin-dUTPs and the amount of input DNA in standard HSE reaction (300ng) about 1% of the initially used DNA amount could be labeled unspecific (detailed information cannot be provided due to confidentiality restrictions). Concentration measurements of the total genomic DNA recovered after HSE based on 300 ng DNA input was determined to be about 100 pg/µl in 50 µl elution volume, would represent even 1.6% of unspecific extraction. It is unclear however how much of this DNA indeed gets captured based on specific and unspecific labeling by biotin-dUTPs or other mechanisms, such as through unspecific (i.e. non-biotin-based) binding to beads, crosshybridization, or physical entanglement to targeted strands.

One reason for any biotin-labeling occurring randomly could be based on any unspecific binding of the probe within the genome, followed by extension. Another reason could be a random extension of any available 3'-termini inside the genome.

Conducting HSE reactions in the absence of probes showed clear evidence that biotin incorporation does not only occur during the extension of the probe but must happen randomly throughout the genome. Surprisingly these control reactions often showed considerable background signals and complete profile detection. Moreover, some results indicated that the obtained background seemed to be influenced by the DNA quality or age (data not shown). In this study no systematic analysis of HSE with different DNA quality was carried out but first results indicate that HSE with older DNA mixes result in more background signals and less separation success for both HSE with or without probes (data not shown).

For the synthesis of a new DNA strand the *Taq* polymerase requires, next to template DNA and magnesium ions, a free 3'-OH group at the desoxyribose sugar. Therefore, the most obvious DNS damages which would affect HSE strongly would be an increased availability of randomly distributed free 3'-OH groups in the entire genome. Because *Taq* polymerase possesses a 5'-nuclease domain in addition to its polymerase domain, even small single strand breaks (nicks) or gaps could serve as a starting point for new strand synthesis and therefore also for unspecific biotin incorporation.

Although *Taq* polymerase is not the enzyme of choice for nicktranslation assays, it has been shown to remove nucleotides from the 5'-end of a DNA chain and facilitate nick translational DNA synthesis [186-188]. The well-known TaqMan assay uses the 5' to 3'-exonuclease function of *Taq* to remove the fluorogenic probe during the extension phase of the PCR [189]. Furthermore, Lyamichev and co-workers hypothesized that 5'-nuclease activity may cleave any unpaired 5'-arm of overhanging

fragments and thus release fragments up to 200 bp long with 3'-OH groups [190, 191]. However, the formation of 3'-OH is mainly known to be catalyzed in enzymatic reactions due to the activity of several cellular enzymes during the process of DNA repair, recombination and apoptosis [192]. Such cellular DNA nicks would remain after DNA extraction and could serve as additional start sites for new strand synthesis during HSE via nick translation.

Although non enzymatic reactions or spontaneous DNA degradation occur more slowly, they also might be sufficient to introduce a rapid loss of HSE efficiency. Aqueous solutions of DNA are sensitive to depurination, depyrimidation, deamination and hydrolytic cleavage [193-196]. The glycosidic base-sugar bond is the most susceptible bond to cleavage within the polynucleotide strand. It has been shown that depurination of DNA in solution took place at a significant rate and is followed by a primarily ß-elimination [197, 198]. In a medium with ionic concentrations similar to intracellular compositions the apurinic site has an estimated half life of between 288 and 335 hours [199]. Depyrimidation of natural bases also occurs in DNA but at rates 20-fold lower than for depurination at neutral pH [200]. DNA bases and desoxyribose sugars are also susceptible to degradation by free radical oxidation, leading to the production of oxidized bases and strand breaks.

The most relevant sources of the formation of reactive oxygen species are metal-catalyzed reductions of oxygen [201, 202]. Hereby, small metal impurities can significantly compromise DNA degradation, and the storage of DNA without further chelators increases the formation of reactive oxygen species [194, 203-205]. Therefore it can be concluded that also during the storage of DNA solutions degradation occurs through the two major pathways of depurination/ß-elimination and free radical oxidation. Hereby, the latter would even be assisted by the storage of DNA in only water and without chelators (such as TE buffer). Such degradation prior to HSE hybridization reaction could enhance DNA damage or facilitate strand breaks at apurin sites.

Another further important factor for DNA damage and storage is the temperature. Although, DNA solutions are stored at 4°C, DNA is exposed to high temperature during denaturation step of the HSE. Heat-induced DNA lesions have been described for DNA strand breaks [206-208], the hydrolysis of glycosyl bonds, (predominantly with the release of purine base [198]) and for the formation of reactive oxygen species [209].

However, the majority of the DNA degradation described in the literature does not result in 3'-OH groups. For example, common products of strand breakage are 3'- and 5'-termini with phosphate groups. Nevertheless, considering that 5' to 3'-exonuclease activity of the *Taq* polymerase is low, also additional strand breaks could increase the availability of existing free 3'-OH. Hereby, nicks and small gaps which contain a 3'-OH could be more accessible for the *Taq* polymerase with an increasing number of strand breaks which were generated by DNA degradation during storage. After the

denaturation and the random reassociation during HSE, an additional number of ss and ds breaks could increase the formation of several single stranded parts inside the rehybridized DNA double strands [192]. Therefore, more 3'-OH groups of nicks and gaps could be better accessible by the *Taq* enzyme, thus increasing unspecific biotin-labeling during HSE.

However, it cannot be excluded that 3'-OH formation also occurs to a very minor extent during non enzymatic DNA degradation [210]. Studies of heat induced ß-elimination at apurin sites showed to a small extent also the formation of desoxyribose with a 3'-OH group [211]. Also, metal-activated hydrolysis of phosphor diester bonds has been reported to create 3'-OH groups [212]. Therefore, it can be assumed that to some minor extent also new strand breaks with free 3'-OH groups are formed randomly, which then interfere in specific HSE reactions.

Next to the DNA quality also probe specificity could influence HSE background. Probe specificity has been observed to be the most important factor for HSE efficiency (chapter 3.4.1). Unspecific probe binding near the target site reduced the separation success dramatically (chapter 3.4.2, Figure 3-9). Furthermore, it must be considered that unspecific extensions of the hybridized probe can occur randomly in the entire genome. Those obviously depend on the hit frequency of the probe. Finally, it has to be mentioned that background signals could be also influenced by the competition of primers of the Y-filer multiplex reaction. As described before the minor component in a mixed DNA sample can be detected up to a ratio of 1:20 (chapter 2.14.3). In the case of a high concentrated major component (> 1ng DNA input) the minor DNA component can fail detection although it expose sufficient DNA material. In the case of a HSE reaction this would mean that the amplification of the strongly enriched Y-STR loci would be favored, whereas underrepresented minor components of the non-targeted alleles could fail detection. Indications of this effect show the Y-chromosomal and autosomal multiplex PCR analysis (Y-filer and NGM) of successful HSE reactions. Hereby, the non-targeted loci of the Y-filer PCR showed drop outs, while the NGM multiplex PCR showed complete profiles with peak areas between 3000 and 12000 (Figure 4-3).

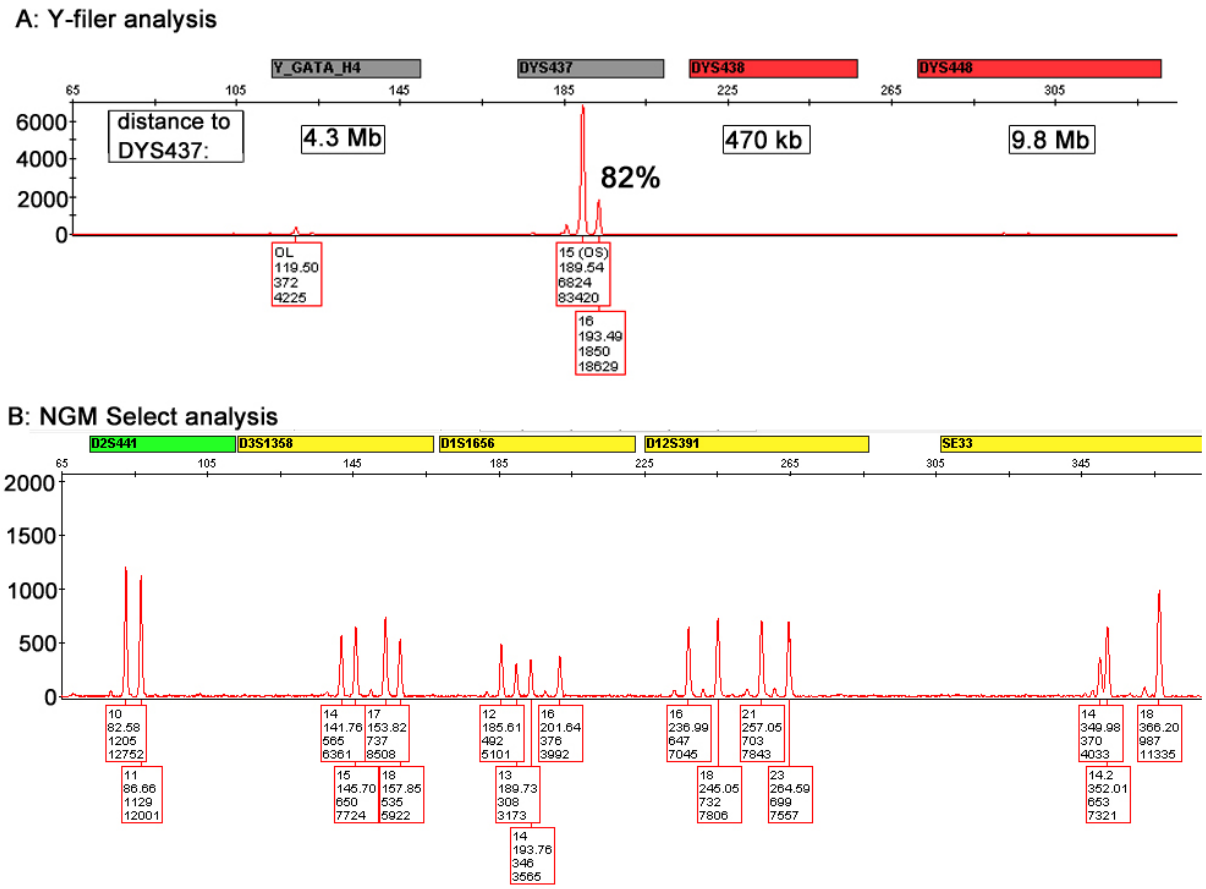


Figure 4-3: Comparison of background detection with the Y-filer and NGM multiplex kit from same HSE sample.
The same HSE sample was analyzed with (A) the Y-filer or (B) the NGM Select multiplex kit. HSE was carried out after the protocol for multiplex I with 300 ng of DNA mix. The figure shows only the red channel of the electropherogram.

4.2.5 Strategies and prospects for HSE in forensic investigations

The application of the HSE technique for the separation of forensic male mixtures holds new challenges, such as the design of HSE for unknown samples and typically DNA samples of poor quality and very limited quantity (Figure 4-4).

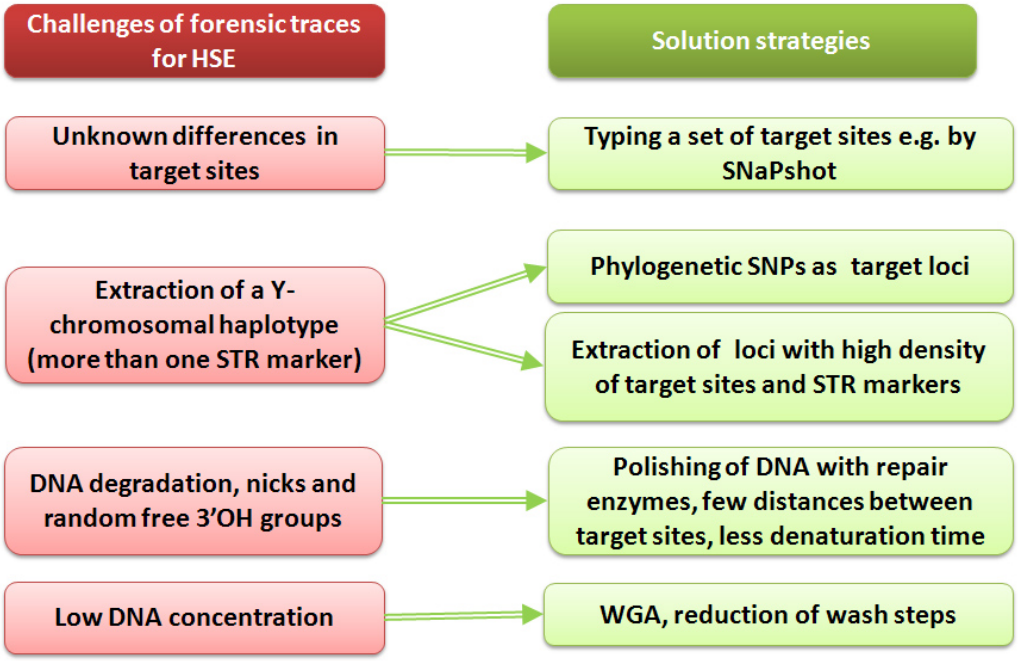


Figure 4-4: Typical issues and possible solutions for the application of HSE on forensic mixtures.

4.2.5.1 Strategies for the extraction of Y-STR haplotypes containing several STR markers

One of the most important requirements for the successful use of HSE in forensics is the simultaneous separation of several linked STR marker of one contributor, since only a combination of markers can be used for individual identification. The results in this study showed the simultaneously extraction of six Y-STR markers from one contributor through the combination of five probes in a multiplex HSE. These five probes were designed for SNPs that were informative based on known ancestry (also called phylogenetic SNPs[5]) as target loci. Hereby, the used SNPs define the haplogroups of the two contributors, named "I" or "R" (chapter 1.5). Because all known Y-chromosomal phylogenetic markers are organized in an ancestry-tree structure and are clearly assigned to defined branches, the information about the haplogroup composition of a DNA mixture allows one to select several target loci along the Y-chromosome.

[5] except rs13304202

Phylogenetic SNPs therefore present a very elegant way to combine several extraction loci into one haplotype. However, an extension of the here described multiplex HSE I and II to all loci included in the Y-filer kit for the used DNA mix would only be possible with the use of further non-phylogenetic SNPs, since no other known SNPs of the haplogroups “R” or “I” are located near additional Y-filer STR markers. This example illustrates the two difficulties when HSE is used for the separation of unknown male haplotypes.

First, after the analysis of a mixed Y-chromosomal profile with the Y-filer, haplogroups can be only predicted. Therefore only little information is given about possible SNP-differences and possible target loci are not known and first have to be analyzed. This can be achieved by the set up of multiplexed SNP analysis, such as a SNaPshot assay, which has been shown to provide excellent results for forensic traces and requires only a small sample volume [22, 213, 214].

The second problem is the availability of target loci near suitable Y-chromosomal STR markers. Contributors of one mixture could belong to the same haplogroup and complicates the combination of different extraction sites, since they do not differ in phylogenetic SNPs. In addition, the random combination of contributors will also show different combination of available target loci, since contributors naturally will differ in respective loci and require a different set of available extraction sites.

At present, there are about 600 Y-chromosomal ancestry-informative markers which together comprise a phylogenetic tree with 311 haplogroups [142]. Although a number of these phylogenetic SNPs are located near Y-chromosomal STR markers which are included in currently available STR profiling kits like the AmpFLSTR® Yfiler, it would be necessary also to use additional Y-STR markers. For example, an extraction of all currently existing Y-filer STRs would not be feasible because it would be required that the haplogroup combination of the mixture shows at least one difference in their phylogenetic branches near every STR marker typed by the Y-filer. Adding non-phylogenetic SNPs to the multiplex HSE would be only possible when the contributors are known. However, if one does not stay with the limited number of forensic standard Y-STRs and use the entire potential of available Y-chromosome STR markers, the number of possible combinations of target sites and therefore the number of simultaneously target STR loci increases enormously. In order to get a first impression of the potential of the Y-chromosome for HSE design, 390 STR markers (published by Hanson and Ballyntyne, 2006 [215]), plus 41 additional markers from the family tree database plus 577 published ancestry-informative SNP markers (published by Karafet *et al.* 2008 [142]) were compared according to their distances to each other. This comparison showed that 24 Y-STR systems were found with distances of less than 1 kb to nearest phylogenetic SNP, 118 Y-STRs with distances between 1kb and 25kb, and an additional 57 Y-STRs with distances between 25 kb and 50 kb.

These 199 different STR markers are located in the vicinity of 347 different phylogenetic SNP markers, which define for 191 haplogroups. Furthermore, in 1047 instances a Y-STR marker was found to be close to a SNP (< 50 kb), which means that on average every Y-STR marker is near five different phylogenetic SNPs (Appendix-Table 7-4, Appendix-Figure 7-3).

In order to demonstrate the usefulness of the phylogenetic tree for HSE design, a mixture with the most widely represented haplogroups in Europe, "R" and "I" [142, 216] and their most frequent sub-branches, R1b1b1a1a-U106 or R1b1b1a1b-S116/P312 [217-219] and I1a, were chosen as an example [220]. The comparison of all possible combinations of Y-STRs and SNP markers showed 54 potential target sites for 108 different STR markers (Figure 4-5). Hereby, 21 of the STR markers showed two, nine STR markers showed three, and two STR markers even showed four possible extraction loci. For example, looking at Figure 4-5 the STR marker DYS66 appears in the near of the three possible extraction sites M258, M253 and M173. Furthermore, 28 of the 108 STR markers are less than 10 kb away from the next SNP. Although some of the STR markers might be not suitable for forensic purposes, this comparison shows the potential power of the HSE design and the possibility of the extraction of haplotypes that include a substantial number of STR markers. In addition, the continuously growing number of phylogenetic SNPs will further extend the combination of STRs and facilitate HSE probe design. This already became apparent during the comparison of the two haplotype branches "R" and "I". Since the last release of the phylogenetic tree in 2008, the use of more recent publications extends the comparison by four additional SNPs and three STR markers (Figure 4-5).[6]

Another viable strategy for the combination of several Y-chromosomal STR markers would be the use of individual, lineage-informative SNPs as extraction loci that are located in highly variable regions with several nearby Y-STR markers (Figure 4-6). In that case HSE can be carried out with only one probe resulting in the separation of several markers. Finally, a combination of the two strategies - the use of phylogenetic and lineage-informative SNPs - would be possible if both are in close proximity. Both strategies however include the extension of currently available standard forensic markers.

Moreover, this also permits the development of a standardized extracted haplotype set, because new combinations of individuals in a mixture may require the use of different extraction loci, thus resulting in different separated haplotypes. The use of new markers would not be problematic for the accurate comparison and exclusion of victims or suspects, but the HSE results with such new markers could not be used for current database searches after ancestry origin at the present, since this database still contains only a small set of markers. However, when considering the most recent

[6] Phylogenetic markers should only be used if their position in the tree is confirmed, since multiple positions in the tree (such as reported for the SNP P25) would lead to incorrect haplotype correlations [221].

developments and efforts in the use of next generation sequencing for the typing of forensic markers [222-224], the future promises rapidly growing reference databases of an enormous set of suitable markers.

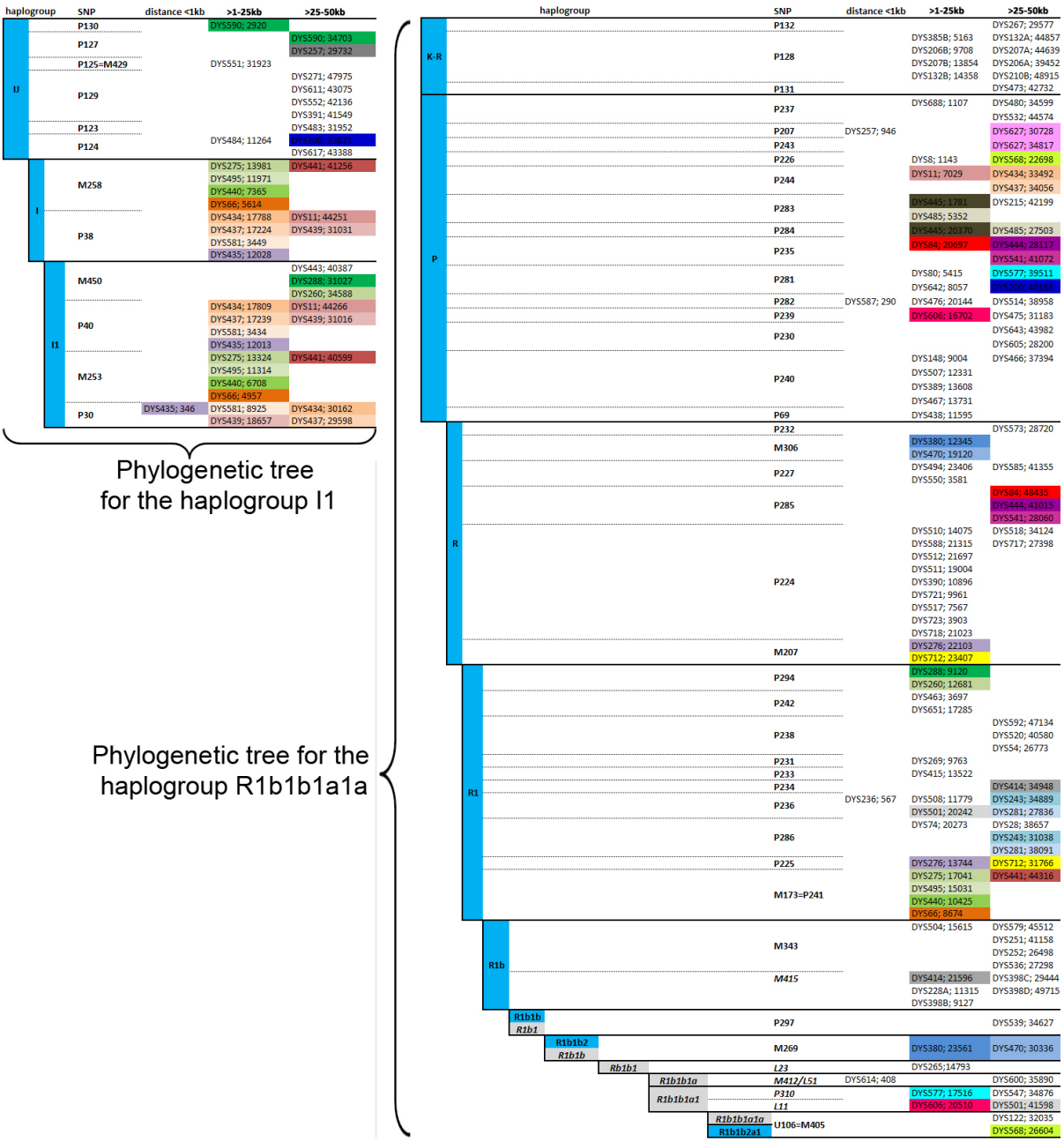


Figure 4-5: Phylogenetic SNPs with nearby Y-STR markers for two high frequent European haplogroups.
A male DNA mixture, from two individuals with the common haplogroups I1 and R1b1b1a1a containing 54 different phylogenetic SNPs or HSE extraction loci. The basic phylogenetic tree structure is taken from the last publication of the YCC in 2008 (blue). The recent publication of new R1b markers are added as grey boxes with italic letters; L514, L11, M412 [218], P310 [217]. Nearby Y-STR markers are grouped per distance to the nearest phylogenetic SNP. STR markers which appear twice or three times are labeled with the same color.

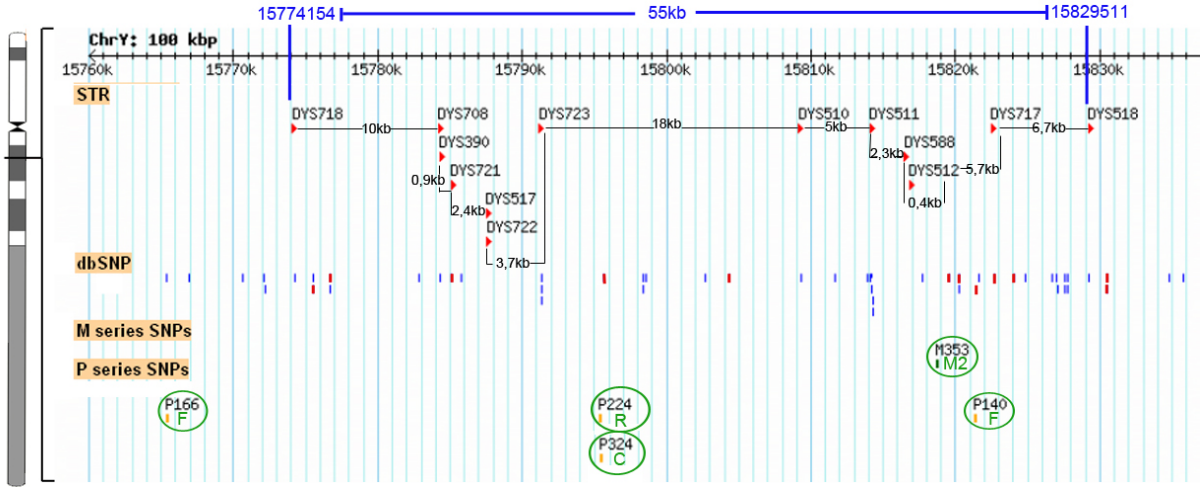


Figure 4-6: Position of 11 neighboring Y-STR systems, spanning a distance of 55 kb.
The map is taken from the Family Tree Database and shows the position of Y-STR systems relative to phylogenetic SNPs (M and P series in green circles) and SNPs from the NCBI database (dbSNPs) on a 55kb segment of the Yq11.221 region. Numbers between DYS markers indicate the distance to an adjacent marker. Note that DYS708/DYS390 and DYS517/DYS722 are redundantly designated markers. Small blue lines indicate NCBI SNP entries with no information or very small sample sizes. Small red lines indicate NCBI SNP entries with several independent submission numbers or population studies. Green letters for phylogenetic SNPs indicate branches of the genetic tree.

4.2.5.2 Strategies to overcome poor quality and quantity of forensic samples before HSE

The best templates for HSE extractions are freshly prepared DNA solutions with long and intact genomic fragments. HSE from DNA mixes with low quality may show a reduction in the separation success which could be caused by an increased number of single strand or double strand breaks of the genome. As a result the number of accessible 3'-OH sites rises, which in turn triggers unspecific extension by the polymerase (see chapter 4.2.4.2). In this study, male DNA mixtures were prepared fresh from frozen blood and therefore display high quality. However, most forensic samples typically have been exposed to far less optimal conditions for unknown time and therefore display a reduced DNA quality which can disturb the HSE extraction. For example, many forensic samples are exposed to light, water or soil, which promotes DNA degradation. Although many of these samples provide sufficient amount of DNA for forensic typing and display only small degrees of degradation, HSE might fail due to its high sensitivity towards strand breaks. Furthermore, forensic samples cover various types of human tissues, e.g. drops of blood, saliva, sperm or other. It has been shown that the cellular content of 3'-OH residues varies between different tissues and was shown to be higher for example in muscle cells or salivary glands [192, 225, 226]. Therefore, HSE efficiency should be tested with DNA mixtures from different tissues.

One promising strategy to overcome this restriction of HSE to only high molecular weight DNA would be the elimination of available 3'-OH residues in the genome before extraction. Free 3'-OH groups from random nicks and gaps can be closed by the combined activity of a polymerase and a ligase

enzyme before HSE. Repair enzymes like T4 ligase and *Escherichia coli* polymerase I were already been tested for the re-establishment of ancient DNA [227, 228]. Although this method showed only a partial recovery of informative markers from ancient DNA, it can be expected to have a higher success rate for 3'-OH repair of damaged, modern DNA.

Furthermore, in order to prevent degradation of the sample during preparation and analysis, DNA samples should be stored with metal chelators like EDTA and HSE should be used with reduced denaturation time and /or reduced temperature as already explained in chapter 4.2.4.2.

In addition to DNA repair before HSE, it might be also promising to test polymerases without 5' to 3'-exonuclease activity. Table 4-1 shows a selection of commercially available polymerases without any exonuclease function, and therefore without the ability of unspecific extension of nicks. From this excerpt of enzymes only the KlenTaq or the VentR® (exo-) would be suitable for HSE. The KF (exo-) has to be excluded because it is not thermostable. The polymerase IV has been listed because of its missing ability for strand displacement. On one hand the ability of the polymerase for strand displacement could contribute to an unspecific extension of free 3'-OH. On the other hand it is not clear if strand displacement might be required or advantageous also for biotin-labeling and successful extraction during HSE. It is known that the strand separation activity resides in the polymerase domain [229, 230]. Therefore the ability for strand displacement becomes difficult to suppress. The polymerase IV of *sulfolobus* belongs to the pol (Y) family with low fidelity and therefore it is not able to discriminate mismatches [231]. However, the polymerases KlenTaq or VentR® (exo-), present some promising candidates to reduce background formation during HSE.

Table 4-1: Alternative polymerases for HSE without exonuclease activity for use in HSE.

polymerase	5' to 3' exonuclease (nicktranslation)	3' to 5' exonuclease (proofreading)	strand displacement	thermo-stable	error rate ($x10^{-6}$)
Taq	+	-	+ [b]	++	20-110 [1, 30, 234, 235]
Stoffel fragment / KlenTaq [a]	-	-		++	51 [234]
KF (exo-)	-	-	+++ [229, 230]	-	100 [236]
VentR® (exo-)	-	-	+++	+++	190 [237]
Sulfolobus DNA Polymerase IV	-	-	-	+	

Characteristics have been partly taken from product guides from New England BioLabs® Inc. and Applied Biosystems. a) Characteristics of KlenTaq (BioLabs) and Stoffel (Applied Biosystems) are presented together because they are assumed to be similar [232, 233]. b) *Taq* polymerase destroys displaced strand by its 5'-nuclease activity.

Another aspect of DNA quality - next to its degradation - is its quantity or DNA concentration. Forensic samples vary highly from few picograms to some nanograms per microliter. Tests of

different multiplex HSE reactions with different amounts of DNA input showed a decrease of the separation success beginning from 50 ng to a minimum of 5 ng of total DNA in a mixture. Therefore, a number of forensic samples would not deliver sufficient DNA material for HSE analysis.

The amount of polymerase and biotin-dUTP was observed to influence the intensity of extracted background signals. Simulation of probe hybridization with higher probe concentration by the program Visual Omp™ showed an increase of duplex concentration. Therefore, one strategy to improve HSE extraction efficiency was tested by the set up of HSE reactions with low DNA amount but higher concentration of polymerase, biotin-dUTP or probe. First tests however did not indicate an improved extraction efficiency, which stopped us from carrying out further tests (data not shown). Another strategy which could overcome the limited DNA concentration of certain forensic samples for HSE is whole-genome amplification (WGA). Minimal amounts of DNA for reliable amplification of forensic markers by WGA have been shown to be around 1 ng. The increase in WGA error rates when smaller amounts of input DNA are used might be compensated for through the use of pooled WGA samples [238]. Therefore, WGA could be used to amplify the available genomic DNA of the sample before HSE to deliver sufficient DNA material for the extraction of both contributors. Furthermore, it has been shown that during the HSE reaction DNA, which bound unspecific to the beads, could be removed by sufficient washing. However, in HSE with only few DNA input it could be favorable to reduce also the washing steps during HSE. In addition the unspecific binding of DNA to the beads could be decreased by using beads, which are blocked with a DNA component such as salmon sperm DNA.

5 Conclusions

At the beginning of this work an intensive SNP search in public database was carried out, in order to find suitable extraction sites for the design of HSE probes near Y-STR markers. Hereby several SNPs from databases have been analyzed and gave rise for a build of guidelines for excluding false entries.

SNP database – best practices to exclude false entries

- Public SNP entries should always be considered carefully; the primary errors that give rise to false SNP entries (sequencing or assembly errors, paralogous sequence variants (PSV)s, and private alleles) should systematically be eliminated. Therefore the following guidelines are recommended:
  I. Search for SNPs on different available databases: look for SNPs with exact mapping positions and criteria of interest (e.g., the type of SNP, its position in introns or exons, its clinical associations)
  II. Exclude PSVs by confirming a low SNP density of the target region and the absence of BLAST matches in paralogous or duplicated regions
  III. Check the validation status of each SNP with its confirmation status, allele frequency and all available information, including the discovery strategy, sample size, characteristics of the method used to identify the variant (computer-generated or experimental, big or small scale), and any special features of the sequence loci

- This strategy maximizes the exclusion of false entries, even in case of little or wrong database information, and provides an efficient, cost-effective method for checking SNP validity. However, uncommitted SNPs are difficult to detect. Therefore an independent internal analysis still remains the best check and each scientist must decide whether to investigate further, based on their own cost/benefit assessment.

In the main part of this study the haplotype-specific extraction was evaluated for the separation of a Y-chromosomal haplotype of a male DNA mixture. Therefore the method was investigated by

addressing several aspects: the HSE protocol steps, the HSE hybridization buffer, HSE probe specificity and HSE unspecific background. All observed and simulated correlations between different parameters on the separation success of male DNA mixtures are summarized in the schematic overview in Figure 5-1.

Separation of Y-chromosomal haplotypes:

- Separation rate does not improve with decreased dNTPs or changed biotin-dUTP concentrations.
- Reduction of the polymerase activity increases the HSE success and therefore the mismatch discrimination.
- Probe length influences the separation effect dramatically. Influence of type of mismatch on probe specificity could not be observed.
- Mismatch discrimination by polymerase and therefore probe specificity and separation success depends on the difference between duplex concentration of matched and mismatched probe. High HSE success correlates with maximum $\Delta con_{M\text{-}MM}$.

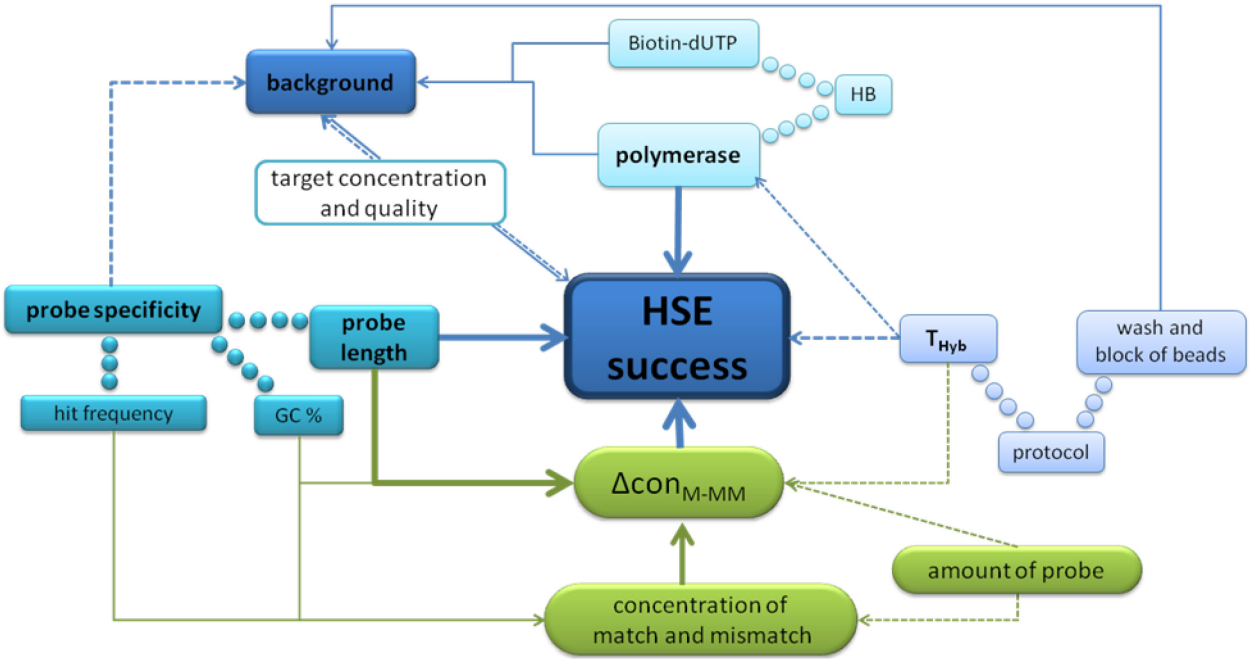


Figure 5-1: Summary of parameters which are affecting HSE success.
The HSE technique includes the three major parts; HSE protocol, hybridization buffer and the probe. Different shades of blue boxes describe groups of parameters which were tested experimentally. Green boxes present parameters whose influence was simulated with the program Visual Omp™. In this study the tested and obtained influences and correlations between parameters are indicated with continuous arrows. Influences which were only simulated or tested in few experiments are given in dotted arrows.

- Probe specificity has to be considered to depend also on the hybridization temperature, as it also indicated in the concentration versus temperature diagrams. However, changes of the hybridization temperature would also influence the polymerase activity, which in turn influences HSE success.
- Probe length with maximal Δcon depends on G/C contents. Higher G/C contents shifts $\Delta con_{M\text{-}MM}$ to smaller probe lengths and vice versa.
- Concentration of match and mismatch are in turn influenced by probe parameters like hit frequency of the probe in the genome and G/C contents and could also directly affect separation success as it was demonstrated for the probe S6:4204FG.
- Best probe length with maximal Δcon can be predicted by the help of the primer design program Visual Omp™ to reduce experimental effort for evaluation work of probe design.
- Simulated higher concentrations of probe in HSE reactions also increase the concentration of hybridized probe, which could be observed in experiments with higher probe concentrations. Furthermore, simulation shows that this also shifts $\Delta con_{M\text{-}MM}$ to different probe length as it is proposed for the probe set P38.
- Maximum distance between target locus and analyzed marker is 50 kb, reflecting the maximum length of extracted fragments.
- Haplotypes of several STR markers can be extracted by multiplexed HSE and the use of phylogenetic SNPs.
- 50ng to 5 ng of genomic DNA is the minimal amount at which separation still occurs.
- During HSE DNA binds unspecific at streptavidin coated magnetic beads. Unspecific bound DNA can be removed efficiently by wash steps.
- Background occurs due to unspecific biotinylation during HSE. Different concentration of *Taq* polymerase and biotin-dUTP effect background signals. Furthermore unspecific binding of the probe can lead to additional background.

6 Prospects

The results of this study uncovered the complex correlation between the different parameters of the HSE approach and the separation success. This new knowledge can be used in future work for further HSE applications, especially in the field of forensic genetic. In this context it is aimed to test the HSE application also after following issues:

- Test of different tissues: Forensic traces arise not only from blood but also from a number of different tissues. In a first test HSE should be carried out with DNA mixtures from different tissues such as salvia or sperm, which deliver in general high DNA concentration. In addition HSE can be tested with set up mixed forensic traces such as cigarettes.
- Test of HSE from DNA samples with less quality: Forensic DNA samples often show degradation and reduced concentration. DNA samples with reduced quality should be tested in HSE in combination with strategies, which aim to overcome the limitation of DNA degradation and low concentration:
    - A ligase assay could help to remove free 3'-OH residues from degraded DNA samples.
    - HSE with reduced wash steps and blocked beads could improve the yield of extracted fragments without increasing the unspecific background.
    - HSE with reduced denaturation temperature and/or time could be beneficial for DNA with low quality.
- Establishing of an approach determining the concentration of an extracted allele: For example, the design of a real-time PCR approach with TaqMan probes specific for extracted and non extracted targets would allow quantification also for a extracted marker. It also would give more information about the loss of DNA material during HSE.
- Design of new multiplex HSE reactions for an extended spectrum of mixed samples: The two different strategies for the design of multiplex HSE, the use of phylogenetic SNPs or the use of SNPs in highly variable regions, should be applied in order to extend the number of separated markers and also to broaden the number of DNA sample combinations.
- On further interesting forensic application would be the use of HSE to separate mt-DNA from DNA mixtures.

7 APPENDIX

7.1 Internet links and web tools

Database searches for Y-chromosomal SNPs and STR systems

- NCBI SNP database (www.ncbi.nlm.nih.gov/projects/SNP/)[239]
- New attributes for NCBI SNP database: (www.//ncbi.nlm.nih.gov/projects/SNP/docs/rs_attributes.html)
- Mammalian Genotyping Service (www.marshfieldclinic.org/mgs/) [171]
- Family Tree DNA (http://ymap.ftdna.com)
- International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/)[240]
- UCSC Genome Bioinformatics (http://genome.ucsc.edu/index.html) [241]
- Human Genome Segmental Duplication Database (http://projects.tcag.ca/humandup/), (Cheung etal. 2003)
- Human Organizes Whole genome Database (http://howdy.biosciencedbc.jp/HOWDYCL/top.pl) [242]

Database searches for HLA

- IMGT/HLA database (http://www.ebi.ac.uk/imgt/hla)
- IPD/MHC database (http://www.ebi.ac.uk/ipd/mhc)

Search for Y –chromosomal Haplotypes and Haplogroupes

- YHRD database (www.yhrd.org)
- US Y-STR Database (http://www.usystrdatabase.org)
- Athey's Haplogroup predictor (http://www.hprg.com/hapest5/hapest5a/hapest5.htm)

- ***Genealogy Y-STR haplotype databases***
- Ysearch (http://www.ysearch.org)
- Sorenson Molecular Genealogy Foundation (http://www.smgf.org)

Forensic databases

- *DNA Analyse Datei* (http://www.bka.de)

Forensic communities

- Scientific Working Group on DNA Analysis Methods (http://www.swgdam.org)
- European Network of Forensic Science Institute (http://www.enfsi.eu)
- European DNA Profiling Group (http://www.isfg.org/EDNAP)
- German DNA Profiling Group (http://www.gednap.de)

- ***Probe and Primer Design***
- Short Tandem Repeat DNA http://www.cstl.nist.gov/div831/strbase/
- NCBI BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi)
- BCM Search Launcher (http://searchlauncher.bcm.tmc.edu/seq-util/readseq.html)
- OligoAnalyzer1.1.2 (http://oligo-analyzer.software.informer.com)
- Oligo Explorer1.2 (http://www.genelink.com)

Sequence Alignments

- ClustalW2 - Multiple Sequence Alignment (http://www.ebi.ac.uk/Tools/msa/clustalw2/)

7.2 Abbreviations

°C, deg	degrees Celsius
6-FAM	6-carbocyfluorescein
Δcon	concentration difference between match and mismatched probe
ΔG, ΔH, ΔS, ΔT	change of free enthalpy, of enthalpy, of entropy, of temperature
A	adenosine
ABI plate sysem	plate set, which consists of 96 septa well, plate cover, base and retainer from Applied Biosystems
ASA / ASP	allele specific amplification / allele specific PCR
ASPE/PEX / PEXT	allele specific extension / primer extension
ASER	allele specific extension reaction
AMELY /X	amelogenin Y/X
AMP	adenosine monophosphate
ATP	adenosine-5'-triphosphate
AZF	azoospermia factor
BCA	bureau of criminal apprehension
BDT	Big Dye terminator
BLAST	basic local alignment search tool
bp, kb, Mb	base pair, kilobase pairs, megabase pairs
C	cytidine
CCD	charge-coupled device
CE	capillary electrophoresis
C_T	treshold cycle
Da	dalton
DAD	DNA *Analyse Datei*
dATP, dATPαS	2'Desoxyadenosine-5'-Triphosphate, deoxyadenosine a-thiotriphosphate
dCTP	2'Desoxycytidine-5'-Triphosphate
ddNTPs	dideoczribonucleoside triphosphates
dGTP	2'Desoxyguanosine-5'-Triphosphate
dichloro[R6G]	dichloro[Rhodamine 6 G]
dichloro[ROX]	dichloro[6-carboxyl-X-Rhodamine]
dichrloro[TAMRA]	dichlroro[carboxytetramethylrhodamine]
diclroro[R110]	dichlroro[Rhodamine 110]
DIP	insertion-deletion polymorphism
dITP	2'-Deoxyinosine-5'-Triphosphate
DNA	deoxyribonucleic acid
dNTP	desoxyribonucleosidtriphophate
dRhodamine	dichlororhodamine
dTTP	2'-Desoxythymidine-5'-Triphosphate
dUTP	2'-Deoxyuridine, 5´-Triphosphate
EDNAP	European DNA profiling group
EDTA	ethylenediaminetetraacetic acid
ENSFI	European network of forensic science institutes
ESS	European standard set
ExoSAP-IT	exonuclease I and shrimp alkaline phosphatase
FS-Taq	AmpliTaq DNA polymerase, FS
G	guanosine
GEDNAP	German DNA profiling group
Hae	*Haemophilus aegyptius*
HB	hybridization buffer
HLA	humane leukocyte antigen
HRM	high resolution melting cureve

HSE	haplotype specific extraction
ISFG	International Society of Forensic Genetics
K	equilibrium constant
KF	Klenow Fragment
LCN	low copy number
LNA	locked nucleic acid
LR	likelihood ratio
M	molar mass
MGS	mammalian genotyping service
MHC	major histocompatibility complex
min	minute
ml, µl	milliliter, microliter
mM, µM, fM	millimolar, micromolar, femtomolar
mm, µm, nm	millimeter, micrometer, nanometer
MSA	mobility-shifting analog
MSY	male-specific region
MW	molecular weight
Mya	million years ago
µg, ng, pg	microgram, nanogram, picogram
NCBI	national center for biotechnology information
NN	nearest-neighbour
NRY	non-recombining region of Y-chromosom
nt	nucleotide
PAR	pseudoautosomal region
PCR	polymerase chain reaction
PE	probability of exclusion
pH	*pondus hydrogenii*
PPi	pyrophosphate
probe-set	describes one type of probe (same locus, same orientation and same mismatch position at the 3'
PSV	paralogous sequence variant
RFLP	restriction fragment length polymorphism
RFU	relative fluorescence units
rpm	revolutions per minute
RSE	region specific extraction
rth	recombinant *Thermus thermophilus*
S	degenerated base code for G or C
s, ms	second, millisecond
SBE	single base extension
SNP, dbSNP	single nucleotide polymorphism, short variance database of the NCBI
SPSS / PASW	statistical package for social sciences / predictiv analysis software
SRY	sex-determining region Y
STR	short tandem repeat
SWGDAM	scientific working group on DNA analysis methods
T	thymidine
Ta	annealing temperature
Taq	*Thermus aquaticus*
TBE	TRIS-borate-EDTA-buffer
T_{Hyb}	hybridization temperature
TRIS	2-Amino-2-(hydroxymethyl)-propan-1,3-diol
UCSC	university of california santa cruz
UTP	uridine-5'-triphosphate

VNTRs	variable number of tandem repeats
W	degenerated base code for A or T
WGA	whole genome amplification
YCC	Y chromosome consortium
Y-filer STR	short tandem repeat marker, which is included in the multiplex PCR of the AmpFLSTR® Yfiler
YHRD	Y-chromosome haplotype reference database
Y-STR	short tandem repeat on the Y-chromosome

7.3 References

1. Tindall, K.R. and T.A. Kunkel, **Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase**. Biochemistry, 1988. 27(16): p. 6008-13.

2. Dapprich J., C., M.A., **Method for selectively isolation a nucleic acid**. 2001.

3. Nagy, M., *et al.*, **Haplotype-specific extraction: a universal method to resolve ambiguous genotypes and detect new alleles - demonstrated on HLA-B**. Tissue Antigens, 2007. 69(2): p. 176-80.

4. Dapprich, J., *et al.*, **Identification of a novel HLA-DPB1 allele (DPB1*1902) by haplotype-specific extraction and nucleotide sequencing**. Tissue Antigens, 2007. 69(3): p. 282-4.

5. Dapprich, J., *et al.*, **Identification of a new HLA-B allele (B*1576) by haplotype specific extraction**. Hum Immunol, 2007. 68(5): p. 418-21.

6. Lim, J., *et al.*, **A novel HLA-Cw*0436 allele identified by haplotype-specific extraction and sequence-based typing**. Tissue Antigens, 2009. 73(1): p. 78-9.

7. Lim, S.K., *et al.*, **Variation of 52 new Y-STR loci in the Y Chromosome Consortium worldwide panel of 76 diverse individuals**. Int J Legal Med, 2007. 121(2): p. 124-7.

8. Ulrich, S., *et al.*, **A new HLA-B*44 variant, B*4453, identified by haplotype-specific extraction and sequencing-based typing**. Tissue Antigens, 2008. 71(1): p. 91-2.

9. Ulrich, S., *et al.*, **A new HLA-A*26 variant, A*2637 identified by haplotype-specific extraction and sequencing-based typing**. Tissue Antigens, 2009. 73(3): p. 275-276.

10. Gabriel, A., *et al.*, **Global mapping of transposon location**. PLoS Genet, 2006. 2(12): p. e212.

11. Gupta, T., *et al.*, **Microtubule actin crosslinking factor 1 regulates the Balbiani body and animal-vegetal polarity of the zebrafish oocyte**. PLoS Genet. 6(8): p. e1001073.

12. Visscher, P.M., *et al.*, **Five years of GWAS discovery**. Am J Hum Genet. 90(1): p. 7-24.

13. Alderborn, A., A. Kristofferson, and U. Hammerling, **Determination of single-nucleotide polymorphisms by real-time pyrophosphate DNA sequencing**. Genome Res, 2000. 10(8): p. 1249-58.

14. Cai, Y., *et al.*, **Allele-specific extension allows base-pair neutral homozygotes to be discriminated by high-resolution melting of small amplicons**. Anal Biochem, 2010. 406(1): p. 29-33.

15. Chehab, F.F. and Y.W. Kan, **Detection of specific DNA sequences by fluorescence amplification: a color complementation assay**. Proc Natl Acad Sci U S A, 1989. 86(23): p. 9178-82.

16. Chen, X. and P.Y. Kwok, **Template-directed dye-terminator incorporation (TDI) assay: a homogeneous DNA diagnostic method based on fluorescence resonance energy transfer**. Nucleic Acids Res, 1997. 25(2): p. 347-53.

17. Papanikos, F., *et al.*, **Lateral flow dipstick test for genotyping of 15 beta-globin gene (HBB) mutations with naked-eye detection**. Anal Chim Acta, 2012. 727: p. 61-6.

18. Sokolov, B.P., **Primer extension technique for the detection of single nucleotide in genomic DNA**. Nucleic Acids Res, 1990. 18(12): p. 3671.

19. Newton, C.R., *et al.*, **Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS)**. Nucleic Acids Res, 1989. 17(7): p. 2503-16.

20. Wu, D.Y., *et al.*, **Allele-specific enzymatic amplification of beta-globin genomic DNA for diagnosis of sickle cell anemia**. Proc Natl Acad Sci U S A, 1989. 86(8): p. 2757-60.

21. Kuppuswamy, M.N., *et al.*, **Single nucleotide primer extension to detect genetic diseases: experimental application to hemophilia B (factor IX) and cystic fibrosis genes**. Proc Natl Acad Sci U S A, 1991. 88(4): p. 1143-7.

22. Geppert, M. and L. Roewer, **SNaPshot(R) minisequencing analysis of multiple ancestry-informative Y-SNPs using capillary electrophoresis**. Methods Mol Biol, 2012. 830: p. 127-40.

23. Kornher, J.S. and K.J. Livak, **Mutation detection using nucleotide analogs that alter electrophoretic mobility**. Nucleic Acids Res, 1989. 17(19): p. 7779-84.

24. Konstantou, J., P.C. Ioannou, and T.K. Christopoulos, **Genotyping of single nucleotide polymorphisms by primer extension reaction and a dual-analyte bio/chemiluminometric assay**. Anal Bioanal Chem, 2007. 388(8): p. 1747-54.

25. Chan, K., *et al.*, **A comprehensive HBV array for the detection of HBV mutants and genotype**. Clin Biochem, 2011. 44(14-15): p. 1253-60.

26. Kim, Y., *et al.*, **Crystal structure of Thermus aquaticus DNA polymerase**. Nature, 1995. 376(6541): p. 612-6.

27. Steitz, T.A., **DNA polymerases: structural diversity and common mechanisms**. J Biol Chem, 1999. 274(25): p. 17395-8.

28. Carver, T.E., Jr., R.A. Hochstrasser, and D.P. Millar, **Proofreading DNA: recognition of aberrant DNA termini by the Klenow fragment of DNA polymerase I**. Proc Natl Acad Sci U S A, 1994. 91(22): p. 10670-4.

29. Eom, S.H., J. Wang, and T.A. Steitz, **Structure of Taq polymerase with DNA at the polymerase active site**. Nature, 1996. 382(6588): p. 278-81.

30. Ling, L.L., *et al.*, **Optimization of the polymerase chain reaction with regard to fidelity: modified T7, Taq, and vent DNA polymerases**. PCR Methods Appl, 1991. 1(1): p. 63-9.

31. Huang, Y.P. and J. Ito, **DNA polymerase C of the thermophilic bacterium Thermus aquaticus: classification and phylogenetic analysis of the family C DNA polymerases**. J Mol Evol, 1999. 48(6): p. 756-69.

32. Huang, M.M., N. Arnheim, and M.F. Goodman, **Extension of base mispairs by Taq DNA polymerase: implications for single nucleotide discrimination in PCR**. Nucleic Acids Res, 1992. 20(17): p. 4567-73.

33. Kwok, S., *et al.*, **Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies**. Nucleic Acids Res, 1990. 18(4): p. 999-1005.

34. Sarkar, G., *et al.*, **Characterization of polymerase chain reaction amplification of specific alleles**. Anal Biochem, 1990. 186(1): p. 64-8.

35. Wu, J.H., P.Y. Hong, and W.T. Liu, **Quantitative effects of position and type of single mismatch on single base primer extension**. J Microbiol Methods, 2009. 77(3): p. 267-75.

36. Onodera, K., **Selection for 3'-end triplets for polymerase chain reaction primers**. Methods Mol Biol, 2007. 402: p. 61-74.

37. Ayyadevara, S., J.J. Thaden, and R.J. Shmookler Reis, **Discrimination of primer 3'-nucleotide mismatch by taq DNA polymerase during polymerase chain reaction**. Anal Biochem, 2000. 284(1): p. 11-8.

38. Allawi, H.T. and J. SantaLucia, Jr., **Thermodynamics and NMR of internal G.T mismatches in DNA**. Biochemistry, 1997. 36(34): p. 10581-94.

39. Lucarelli, F., G. Marrazza, and M. Mascini, **Design of an optimal allele-specific oligonucleotide probe for the efficient discrimination of a thermodynamically stable (G x T) mismatch**. Anal Chim Acta, 2007. 603(1): p. 82-6.

40. Bottema, C.D. and S.S. Sommer, **PCR amplification of specific alleles: rapid detection of known mutations and polymorphisms**. Mutat Res, 1993. 288(1): p. 93-102.

41. Day, J.P., *et al.*, **Nucleotide analogs facilitate base conversion with 3' mismatch primers**. Nucleic Acids Res, 1999. 27(8): p. 1810-8.

42. Horne, M.T., D.J. Fish, and A.S. Benight, **Statistical thermodynamics and kinetics of DNA multiplex hybridization reactions**. Biophys J, 2006. 91(11): p. 4133-53.

43. Marmur, J. and P. Doty, **Heterogeneity in deoxyribonucleic acids. I. Dependence on composition of the configurational stability of deoxyribonucleic acids**. Nature, 1959. 183(4673): p. 1427-9.

44. Yakovchuk, P., E. Protozanova, and M.D. Frank-Kamenetskii, **Base-stacking and base-pairing contributions into thermal stability of the DNA double helix**. Nucleic Acids Res, 2006. 34(2): p. 564-74.

45. SantaLucia, J., Jr. and D. Hicks, **The thermodynamics of DNA structural motifs**. Annu Rev Biophys Biomol Struct, 2004. 33: p. 415-40.

46. SantaLucia, J., Jr., **A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics**. Proc Natl Acad Sci U S A, 1998. 95(4): p. 1460-5.

47. SantaLucia, J., Jr., **Physical principles and visual-OMP software for optimal PCR design**. Methods Mol Biol, 2007. 402: p. 3-34.

48. Watkins, N.E., Jr. and J. SantaLucia, Jr., **Nearest-neighbor thermodynamics of deoxyinosine pairs in DNA duplexes**. Nucleic Acids Res, 2005. 33(19): p. 6258-67.

49. Jeffreys, A.J., V. Wilson, and S.L. Thein, **Hypervariable 'minisatellite' regions in human DNA**. Nature, 1985. 314(6006): p. 67-73.

50. Butler, J.M., **Fundamentals of Forensic DNA Typing.** 2009: Academic Press Inc. 520.

51. Jobling, M., M.E. Hurles, and C. Tyler-Smith, **Human Evolutionary Genetics: origins, peoples and disease.** 2004, London/New York: Garland Science Publishing. 523.

52. Buschiazzo, E. and N.J. Gemmell, **The rise, fall and renaissance of microsatellites in eukaryotic genomes**. Bioessays, 2006. 28(10): p. 1040-50.

53. Levinson, G. and G.A. Gutman, **Slipped-strand mispairing: a major mechanism for DNA sequence evolution**. Mol Biol Evol, 1987. 4(3): p. 203-21.

54. Applied-Biosystems, **AmpFlSTR® NGM SElect™ PCR Amplification Kit.** 03 ed. 2011: Life Technologies Corporation. 170.

55. Green, R.L., *et al.*, **Developmental validation of the AmpFLSTR® NGM SElect™ PCR Amplification Kit: A next-generation STR multiplex with the SE33 locus**. Forensic science international. Genetics, 2012.

56. Barber, M.D. and B.H. Parkin, **Sequence analysis and allelic designation of the two short tandem repeat loci D18S51 and D8S1179**. Int J Legal Med, 1996. 109(2): p. 62-5.

57. Zietkiewicz, E., *et al.*, **Current genetic methodologies in the identification of disaster victims and in forensic analysis**. J Appl Genet, 2012. 53(1): p. 41-60.

58. Butler, J.M., M.C. Kline, and A.E. Decker, **Addressing Y-Chromosome Short Tandem Repeat(Y-STR) Allele Nomenclature**. Journal of Genetic Genealogy, 2008. 4(2): p. 125-148.

59. Gill, P., *et al.*, **Considerations from the European DNA profiling group (EDNAP) concerning STR nomenclature**. Forensic Sci Int, 1997. 87(3): p. 185-92.

60. Guichoux, E., *et al.*, **Current trends in microsatellite genotyping**. Mol Ecol Resour, 2011. 11(4): p. 591-611.

61. Butler, J.M., Y. Shen, and B.R. McCord, **The development of reduced size STR amplicons as tools for analysis of degraded DNA**. J Forensic Sci, 2003. 48(5): p. 1054-64.

62. Hartzell, B. and B. McCord, **Effect of divalent metal ions on DNA studied by capillary electrophoresis**. Electrophoresis, 2005. 26(6): p. 1046-56.

63. Magnuson, V.L., *et al.*, **Substrate nucleotide-determined non-templated addition of adenine by *Taq* DNA polymerase: implications for PCR-based genotyping and cloning**. Biotechniques, 1996. 21(4): p. 700-9.

64. Huel, R.L., *et al.*, **Variant alleles, triallelic patterns, and point mutations observed in nuclear short tandem repeat typing of populations in Bosnia and Serbia**. Croat Med J, 2007. 48(4): p. 494-502.

65. Brookes, C., *et al.*, **Characterising stutter in forensic STR multiplexes**. Forensic Sci Int Genet, 2012. 6(1): p. 58-63.

66. Meldgaard, M. and N. Morling, **Detection and quantitative characterization of artificial extra peaks following polymerase chain reaction amplification of 14 short tandem repeat systems used in forensic investigations**. Electrophoresis, 1997. 18(11): p. 1928-35.

67. Walsh, P.S., N.J. Fildes, and R. Reynolds, **Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA**. Nucleic Acids Res, 1996. 24(14): p. 2807-12.

68. Schlotterer, C. and D. Tautz, **Slippage synthesis of simple sequence DNA**. Nucleic Acids Res, 1992. 20(2): p. 211-5.

69. Mitchell, A.A., *et al.*, **Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in**. Forensic Sci Int Genet, 2012. 6(6): p. 749-61.

70. Applied-Biosystems, **AmpFlSTR® Yfiler™ PCR Amplification Kit.** 08 ed. 2006: Applied Biosystems. 234.

71. Clayton, T.M., *et al.*, **Analysis and interpretation of mixed forensic stains using DNA STR profiling**. Forensic Sci Int, 1998. 91(1): p. 55-70.

72. Applied-Biosystems, **AmpFlSTR® Identifiler® PCR Amplification Kit.** 09 ed. 2006: Applied Biosystems. 186.

73. Barbaro, A. and P. Cormaci, **DNA analysis from mixed biological materials**. Forensic Sci Int, 2004. 146 Suppl: p. S123-5.

74. Schneider, P.M., *et al.*, **The German Stain Commission: recommendations for the interpretation of mixed stains**. Int J Legal Med, 2009. 123(1): p. 1-5.

75. Gill, P., *et al.*, **Interpreting simple STR mixtures using allele peak areas**. Forensic Sci Int, 1998. 91(1): p. 41-53.

76. Gill, P., *et al.*, **National recommendations of the Technical UK DNA working group on mixture interpretation for the NDNAD and for court going purposes**. Forensic Sci Int Genet, 2008. 2(1): p. 76-82.

77. Bright, J.A., *et al.*, **A comparison of stochastic variation in mixed and unmixed casework and synthetic samples**. Forensic Sci Int Genet, 2012. 6(2): p. 180-4.

78. Buckleton, J.S., J.M. Curran, and P. Gill, **Towards understanding the effect of uncertainty in the number of contributors to DNA stains**. Forensic Sci Int Genet, 2007. 1(1): p. 20-8.

79. Budowle, B., *et al.*, **Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework**. J Forensic Sci, 2009. 54(4): p. 810-21.

80. Cowell, R.G., S.L. Lauritzen, and J. Mortera, **Identification and separation of DNA mixtures using peak area information**. Forensic Sci Int, 2007. 166(1): p. 28-34.

81. Torres, Y., *et al.*, **DNA mixtures in forensic casework: a 4-year retrospective study**. Forensic Sci Int, 2003. 134(2-3): p. 180-6.

82. Gill, P., *et al.*, **DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures**. Forensic Sci Int, 2006. 160(2-3): p. 90-101.

83. Weir, B.S., *et al.*, **Interpreting DNA mixtures**. J Forensic Sci, 1997. 42(2): p. 213-22.

84. Ladd, C., *et al.*, **Interpretation of complex forensic DNA mixtures**. Croat Med J, 2001. 42(3): p. 244-6.

85. Bekaert, B., *et al.*, **Automating a combined composite-consensus method to generate DNA profiles from low and high template mixture samples**. Forensic Sci Int Genet, 2012. 6(5): p. 588-93.

86. Pfeifer, C.M., *et al.*, **Comparison of different interpretation strategies for low template DNA mixtures**. Forensic Sci Int Genet, 2012. 6(6): p. 716-22.

87. Voskoboinik, L. and A. Darvasi, **Forensic identification of an individual in complex DNA mixtures**. Forensic Sci Int Genet, 2011. 5(5): p. 428-35.

88. Hall, D. and V. Castella, **DIP-STR: A new marker for resolving unbalanced DNA mixtures**. Forensic Science International: Genetics Supplement Series, 2011. 3(1): p. e1-e2.

89. Jobling, M.A. and C. Tyler-Smith, **The human Y chromosome: an evolutionary marker comes of age**. Nat Rev Genet, 2003. 4(8): p. 598-612.

90. Hughes, J.F. and S. Rozen, **Genomics and genetics of human and primate y chromosomes**. Annu Rev Genomics Hum Genet, 2012. 13: p. 83-108.

91. Kuroda-Kawaguchi, T., *et al.*, **The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men**. Nat Genet, 2001. 29(3): p. 279-86.

92. Skaletsky, H., *et al.*, **The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes**. Nature, 2003. 423(6942): p. 825-37.

93. Lahn, B.T. and D.C. Page, **Functional coherence of the human Y chromosome**. Science, 1997. 278(5338): p. 675-80.

94. Lemaitre, C., *et al.*, **Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes**. Genome Biol Evol, 2009. 1: p. 56-66.

95. Mumm, S., *et al.*, **Evolutionary features of the 4-Mb Xq21.3 XY homology region revealed by a map at 60-kb resolution**. Genome Res, 1997. 7(4): p. 307-14.

96. Page, D.C., *et al.*, **Occurrence of a transposition from the X-chromosome long arm to the Y-chromosome short arm during human evolution**. Nature, 1984. 311(5982): p. 119-23.

97. Schwartz, A., *et al.*, **Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination**. Hum Mol Genet, 1998. 7(1): p. 1-11.

98. Nakahori, Y., O. Takenaka, and Y. Nakagome, **A human X-Y homologous region encodes "amelogenin"**. Genomics, 1991. 9(2): p. 264-9.

99. Stephanopoulos, G., M.E. Garefalaki, and K. Lyroudia, **Genes and related proteins involved in amelogenesis imperfecta**. J Dent Res, 2005. 84(12): p. 1117-26.

100. Sullivan, K.M., *et al.*, **A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin**. Biotechniques, 1993. 15(4): p. 636-8, 640-1.

101. Haas-Rochholz, H. and G. Weiler, **Additional primer sets for an amelogenin gene PCR-based DNA-sex test**. Int J Legal Med, 1997. 110(6): p. 312-5.

102. Mitchell, R.J., *et al.*, **An investigation of sequence deletions of amelogenin (AMELY), a Y-chromosome locus commonly used for gender determination**. Ann Hum Biol, 2006. 33(2): p. 227-40.

103. Kayser, M., *et al.*, **Evaluation of Y-chromosomal STRs: a multicenter study**. Int J Legal Med, 1997. 110(3): p. 125-33, 141-9.

104. Krenke, B.E., *et al.*, **Validation of male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex**. Forensic Sci Int, 2005. 151(1): p. 111-24.

105. Promega, **PowerPlex® Y System.** 2003: Promega Corporation. 58.

106. Mulero, J.J., *et al.*, **Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system**. J Forensic Sci, 2006. 51(1): p. 64-75.

107. Davis, C., *et al.*, **Prototype PowerPlex® Y23 System: A concordance study**. Forensic Sci Int Genet, 2012.

108. Promega, **PowerPlex® Y23 System**. 07/12 ed. 2012: Promega Corporation. 75.

109. Willuweit, S. and L. Roewer, **The Y-Chromsome Haplotpye Reference Database - Direction for use.** 33.1 ed. 2010: http://www.yhrd.org/downloads/manual.pdf.

110. Roewer, L., **Y chromosome STR typing in crime casework**. Forensic Sci Med Pathol, 2009. 5(2): p. 77-84.

111. Prinz, M., *et al.*, **Multiplexing of Y chromosome specific STRs and performance for mixed samples**. Forensic Sci Int, 1997. 85(3): p. 209-18.

112. Gill, P., *et al.*, **DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs**. Forensic Sci Int, 2001. 124(1): p. 5-10.

113. Krawczak, M., **Forensic interpretation of haploid DNA mixtures**. International Congress Series, 2006. 1288: p. 477-483.

114. Roewer, L., *et al.*, **A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males**. Forensic Sci Int, 2000. 114(1): p. 31-43.

115. Parson, W. and L. Roewer, **Publication of population data of linearly inherited DNA markers in the International Journal of Legal Medicine**. Int J Legal Med, 2010. 124(5): p. 505-9.

116. Roewer, L. and M. Geppert, **Interpretation guidelines of a standard Y-chromosome STR 17-plex PCR-CE assay for crime casework**. Methods Mol Biol, 2012. 830: p. 43-56.

117. Butler, J.M., **Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers.** Second ed. 2005, New York: Elsevier Academic Press. 688.

118. Wolf, A., *et al.*, **Forensic interpretation of Y-chromosomal DNA mixtures**. Forensic Sci Int, 2005. 152(2-3): p. 209-13.

119. Roewer, L., *et al.*, **Signature of recent historical events in the European Y-chromosomal STR haplotype distribution**. Hum Genet, 2005. 116(4): p. 279-91.

120. Kayser, M., *et al.*, **Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis**. Hum Genet, 2005. 117(5): p. 428-43.

121. Roewer, L., **Populationsgenetik des Y-Chromosoms**. Medizinische Genetik, 2008. 20(3): p. 288-292.

122. Decker, A.E., *et al.*, **The impact of additional Y-STR loci on resolving common haplotypes and closely related individuals**. Forensic Sci Int Genet, 2007. 1(2): p. 215-7.

123. Ballantyne, K.N., *et al.*, **A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages**. Forensic Sci Int Genet, 2012. 6(2): p. 208-18.

124. Ballantyne, K.N., *et al.*, **Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications**. Am J Hum Genet, 2010. 87(3): p. 341-53.

125. Kayser, M. and K.N. **Ballantyne, Broadening the horizon: New Y-STRs for forensic practice, in ISFG Y-chromosome Workshop.** 2009: Buenos Aires, Argentina.

126. Hanson, E.K. and J. Ballantyne, **Population data for 48 'Non-Core' Y chromosome STR loci**. Leg Med (Tokyo), 2007. 9(4): p. 221-31.

127. Rodig, H., *et al.*, **Evaluation of haplotype discrimination capacity of 35 Y-chromosomal short tandem repeat loci**. Forensic Sci Int, 2008. 174(2-3): p. 182-8.

128. Vermeulen, M., *et al.*, **Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms**. Forensic Sci Int Genet, 2009. 3(4): p. 205-13.

129. Budowle, B. and A. van Daal, **Forensically relevant SNP classes**. Biotechniques, 2008. 44(5): p. 603-8, 610.

130. Kidd, K.K., *et al.*, **Expanding data and resources for forensic use of SNPs in individual identification**. Forensic Sci Int Genet, 2012. 6(5): p. 646-52.

131. Pakstis, A.J., *et al.*, **SNPs for a universal individual identification panel**. Hum Genet. 127(3): p. 315-24.

132. Liu, F., *et al.*, **Eye color and the prediction of complex phenotypes from genotypes**. Curr Biol, 2009. 19(5): p. R192-3.

133. Liu, F., *et al.*, **Digital quantification of human eye color highlights genetic association of three new loci**. PLoS Genet, 2010. 6(5): p. e1000934.

134. Walsh, S., *et al.*, **IrisPlex: A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information**. Forensic Sci Int Genet, 2011. 5(3): p. 170-180.

135. Candille, S.I., *et al.*, **Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four European populations**. PLoS One, 2012. 7(10): p. e48294.

136. Pneuman, A., *et al.*, **Verification of eye and skin color predictors in various populations**. Leg Med (Tokyo), 2012. 14(2): p. 78-83.

137. Zubakov, D., *et al.*, **Estimating human age from T-cell DNA rearrangements**. Curr Biol, 2010. 20(22): p. R970-1.

138. Underhill, P.A. and T. Kivisild, **Use of y chromosome and mitochondrial DNA population structure in tracing human migrations**. Annu Rev Genet, 2007. 41: p. 539-64.

139. Consortium, T.Y.C., **A nomenclature system for the tree of human Y-chromosomal binary haplogroups**. Genome Res, 2002. 12(2): p. 339-48.

140. Cruciani, F., *et al.*, **A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa**. Am J Hum Genet, 2011. 88(6): p. 814-8.

141. Ke, Y., *et al.*, **African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes**. Science, 2001. 292(5519): p. 1151-3.

142. Karafet, T.M., *et al.*, **New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree**. Genome Res, 2008. 18(5): p. 830-8.

143. Sachidanandam, R., *et al.*, **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms**. Nature, 2001. 409(6822): p. 928-33.

144. Xu, J.Y., G.B. Xu, and S.L. Chen, **A new method for SNP discovery**. Biotechniques, 2009. 46(3): p. 201-8.

145. Reich, D.E., S.B. Gabriel, and D. Altshuler, **Quality and completeness of SNP databases**. Nat Genet, 2003. 33(4): p. 457-8.

146. Carlson, C.S., *et al.*, **Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans**. Nat Genet, 2003. 33(4): p. 518-21.

147. Mitchell, A.A., *et al.*, **Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns**. Bioinformatics, 2004. 20(7): p. 1022-32.

148. Estivill, X., *et al.*, **Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome**. Hum Mol Genet, 2002. 11(17): p. 1987-95.

149. Fredman, D., *et al.*, **Complex SNP-related sequence variation in segmental genome duplications**. Nat Genet, 2004. 36(8): p. 861-6.

150. Nakken, S., *et al.*, **Large-scale inference of the point mutational spectrum in human segmental duplications**. BMC Genomics, 2009. 10: p. 43.

151. Musumeci, L., *et al.*, **Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies**. Hum Mutat, 2009. 31(1): p. 67-73.

152. Nelson, M.R., *et al.*, **Large-scale validation of single nucleotide polymorphisms in gene regions**. Genome Res, 2004. 14(8): p. 1664-8.

153. Nagy, M., *et al.*, **Optimization and validation of a fully automated silica-coated magnetic beads purification technology in forensics**. Forensic Sci Int, 2005. 152(1): p. 13-22.

154. Whitcombe, D., *et al.*, **Detection of PCR products using self-probing amplicons and fluorescence**. Nat Biotechnol, 1999. 17(8): p. 804-7.

155. Qiagen**, Investigator Quantiplex Handbook.** 03 ed. 2011.

156. Dean, F.B., *et al.*, **Comprehensive human genome amplification using multiple displacement amplification**. Proc Natl Acad Sci U S A, 2002. 99(8): p. 5261-6.

157. Dean, F.B., *et al.*, **Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification**. Genome Res, 2001. 11(6): p. 1095-9.

158. Chien, A., D.B. Edgar, and J.M. Trela, **Deoxyribonucleic acid polymerase from the extreme thermophile *Thermus aquaticus***. J. Bacteriol., 1976. 127(3): p. 1550-1557.

159. Kayser, M., *et al.*, **A comprehensive survey of human Y-chromosomal microsatellites**. Am J Hum Genet, 2004. 74(6): p. 1183-97.

160. Butler, J.M., *et al.*, **Allele frequencies for 27 Y-STR loci with U.S. Caucasian, African American, and Hispanic samples**. Forensic Sci Int, 2006. 156(2-3): p. 250-60.

161. Ronaghi, M., M. Uhlen, and P. Nyren, **A sequencing method based on real-time pyrophosphate**. Science, 1998. 281(5375): p. 363, 365.

162. Ronaghi, M., *et al.*, **Real-time DNA sequencing using detection of pyrophosphate release**. Anal Biochem, 1996. 242(1): p. 84-9.

163. Sanger, F., S. Nicklen, and A.R. Coulson, **DNA sequencing with chain-terminating inhibitors**. Proc Natl Acad Sci U S A, 1977. 74(12): p. 5463-7.

164. Rosenblum, B.B., *et al.*, **New dye-labeled terminators for improved DNA sequencing patterns**. Nucleic Acids Res, 1997. 25(22): p. 4500-4.

165. Parker, L.T., *et al.*, **AmpliTaq DNA polymerase, FS dye-terminator sequencing: analysis of peak height patterns**. Biotechniques, 1996. 21(4): p. 694-9.

166. Dapprich, J., *et al.*, **SNP-specific extraction of haplotype-resolved targeted genomic regions**. Nucleic Acids Res, 2008. 36(15): p. e94.

167. Gross, A.M., *et al.*, **Internal validation of the AmpFlSTR Yfiler amplification kit for use in forensic casework**. J Forensic Sci, 2008. 53(1): p. 125-34.

168. Istrail, S., *et al.*, **Whole-genome shotgun assembly and comparison of human genome assemblies**. Proc Natl Acad Sci U S A, 2004. 101(7): p. 1916-21.

169. Venter, J.C., *et al.*, **The sequence of the human genome**. Science, 2001. 291(5507): p. 1304-51.

170. Mills, R.E., *et al.*, **An initial map of insertion and deletion (INDEL) variation in the human genome**. Genome Res, 2006. 16(9): p. 1182-90.

171. Weber, J.L., *et al.*, **Human diallelic insertion/deletion polymorphisms**. Am J Hum Genet, 2002. 71(4): p. 854-62.

172. Ning, Z., A.J. Cox, and J.C. Mullikin, **SSAHA: a fast search method for large DNA databases**. Genome Res, 2001. 11(10): p. 1725-9.

173. Ross, M.T., *et al.*, **The DNA sequence of the human X chromosome**. Nature, 2005. 434(7031): p. 325-37.

174. Platzer, M., *et al.*, **Sequencing errors or SNPs at splice-acceptor guanines in dbSNP?** Nat Biotechnol, 2006. 24(9): p. 1068-70.

175. Yu, H., *et al.*, **The new Y chromosome DYS718 in Chinese Han population of Shanxi Genetic Polymorphism**. Modern Medicine and Health, 2008. 144(12): p. 1781-1782.

176. Fenn, B.J. and T.M. Herman, **Direct quantitation of biotin-labeled nucleotide analogs in RNA transcripts**. Anal Biochem, 1990. 190(1): p. 78-83.

177. Theissen, G., A. Richter, and N. Lukacs, **Degree of biotinylation in nucleic acids estimated by a gel retardation assay**. Anal Biochem, 1989. 179(1): p. 98-105.

178. Sommer, S.S., A.R. Groszbach, and C.D. Bottema, **PCR amplification of specific alleles (PASA) is a general method for rapidly detecting known single-base changes**. Biotechniques, 1992. 12(1): p. 82-7.

179. Eckert, K.A. and T.A. Kunkel, **High fidelity DNA synthesis by the *Thermus aquaticus* DNA polymerase**. Nucleic Acids Res, 1990. 18(13): p. 3739-44.

180. Fish, D.J., *et al.*, **Multiplex SNP discrimination**. Biophys J, 2007. 92(10): p. L89-91.

181. Zhao, G. and Y. Guan, **Polymerization behavior of Klenow fragment and *Taq* DNA polymerase in short primer extension reactions**. Acta Biochim Biophys Sin (Shanghai), 2010. 42(10): p. 722-8.

182. Arora, A., *et al.*, **Effect of locked nucleic acid (LNA) modification on hybridization kinetics of DNA duplex**. Nucleic Acids Symp Ser (Oxf), 2008(52): p. 417-8.

183. Gaster, J. and A. Marx, **Tuning single nucleotide discrimination in polymerase chain reactions (PCRs): synthesis of primer probes bearing polar 4'-C-modifications and their application in allele-specific PCR**. Chemistry, 2005. 11(6): p. 1861-70.

184. Strerath, M., *et al.*, **Modified oligonucleotides as tools for allele-specific amplification**. Methods Mol Biol, 2007. 402: p. 317-28.

185. You, Y., *et al.*, **Design of LNA probes that improve mismatch discrimination**. Nucleic Acids Research, 2006. 34(8): p. e60.

186. Dudin, G., *et al.*, **Sorting of chromosomes by magnetic separation**. Hum Genet, 1988. 80(2): p. 111-6.

187. Longley, M.J., S.E. Bennett, and D.W. Mosbaugh, **Characterization of the 5â€² to 3â€² exonuclease associated with *Thermus aquaticus* DNA polymerase**. Nucleic Acids Research, 1990. 18(24): p. 7317-7322.

188. Rigby, P.W., *et al.*, **Labeling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I**. J Mol Biol, 1977. 113(1): p. 237-51.

189. Livak, K.J., **Allelic discrimination using fluorogenic probes and the 5' nuclease assay**. Genet Anal, 1999. 14(5-6): p. 143-9.

190. Lyamichev, V., M.A. Brow, and J.E. Dahlberg, **Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases**. Science, 1993. 260(5109): p. 778-83.

191. Lyamichev, V., *et al.*, **Comparison of the 5' nuclease activities of taq DNA polymerase and its isolated nuclease domain**. Proc Natl Acad Sci U S A, 1999. 96(11): p. 6143-8.

192. Basnakian, A.G. and S.J. James, **Quantification of 3'OH DNA breaks by random oligonucleotide-primed synthesis (ROPS) assay**. DNA Cell Biol, 1996. 15(3): p. 255-62.

193. Alaeddini, R., S.J. Walsh, and A. Abbas, **Forensic implications of genetic analyses from degraded DNA--a review**. Forensic Sci Int Genet, 2010. 4(3): p. 148-57.

194. Anchordoquy, T.J. and M.C. Molina, **Preservation of DNA**. Cell Preservation Technology, 2007. 5(4): p. 180-188.

195. Middaugh, C.R., *et al.*, **Analysis of plasmid DNA from a pharmaceutical perspective**. J Pharm Sci, 1998. 87(2): p. 130-46.

196. Pogocki, D. and C. Schoneich, **Chemical stability of nucleic acid-derived drugs**. J Pharm Sci, 2000. 89(4): p. 443-56.

197. Lindahl, T., **Instability and decay of the primary structure of DNA**. Nature, 1993. 362(6422): p. 709-15.

198. Lindahl, T. and B. Nyberg, **Rate of depurination of native deoxyribonucleic acid**. Biochemistry, 1972. 11(19): p. 3610-8.

199. Crine, P. and W.G. Verly, **A study of DNA spontaneous degradation**. Biochim Biophys Acta, 1976. 442(1): p. 50-7.

200. Lindahl, T. and O. Karlstrom, **Heat-induced depyrimidination of deoxyribonucleic acid in neutral solution**. Biochemistry, 1973. 12(25): p. 5151-5154.

201. Borish, E.T., *et al.*, **Cigarette tar causes single-strand breaks in DNA**. Biochem Biophys Res Commun, 1985. 133(2): p. 780-6.

202. Imlay, J.A. and S. Linn, **DNA damage and oxygen radical toxicity**. Science, 1988. 240(4857): p. 1302-9.

203. Graf, E., *et al.*, **Iron-catalyzed hydroxyl radical formation. Stringent requirement for free iron coordination site**. J Biol Chem, 1984. 259(6): p. 3620-4.

204. Lahiri, D.K. and B. Schnabel, **DNA isolation by a rapid method from human blood samples: effects of MgCl2, EDTA, storage time, and temperature on DNA yield and quality**. Biochem Genet, 1993. 31(7-8): p. 321-8.

205. Molina, M.C. and T.J. Anchordoquy, **Metal contaminants promote degradation of lipid/DNA complexes during lyophilization**. Biochim Biophys Acta, 2007. 1768(3): p. 669-77.

206. Douglas, A. and B. Atchison, **Degradation of DNA during the denaturation step of PCR**. PCR Methods Appl, 1993. 3(2): p. 133-4.

207. Eigner, J., H. Boedtker, and G. Michaels, **The thermal degradation of nucleic acids**. Biochim Biophys Acta, 1961. 51: p. 165-8.

208. Zhang, L. and Q. Wu, **Single gene retrieval from thermally degraded DNA**. J Biosci, 2005. 30(5): p. 599-604.

209. Bruskov, V.I., *et al.*, **Heat-induced formation of reactive oxygen species and 8-oxoguanine, a biomarker of damage to DNA**. Nucleic Acids Res, 2002. 30(6): p. 1354-63.

210. Lindahl, T. and A. Andersson, **Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid**. Biochemistry, 1972. 11(19): p. 3618-23.

211. Sugiyama, H., *et al.*, **Chemistry of thermal degradation of abasic sites in DNA. Mechanistic investigation on thermal DNA strand cleavage of alkylated DNA**. Chem Res Toxicol, 1994. 7(5): p. 673-83.

212. Ott, R. and R. Krämer, **DNA hydrolysis by inorganic catalysts**. Applied Microbiology and Biotechnology, 1999. 52(6): p. 761-767.

213. Onofri, V., *et al.*, **Development of multiplex PCRs for evolutionary and forensic applications of 37 human Y chromosome SNPs**. Forensic Sci Int, 2006. 157(1): p. 23-35.

214. Romanini, C., *et al.*, **Typing short amplicon binary polymorphisms: supplementary SNP and Indel genetic information in the analysis of highly degraded skeletal remains**. Forensic Sci Int Genet, 2012. 6(4): p. 469-76.

215. Hanson, E.K. and J. Ballantyne, **Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications**. Leg Med (Tokyo), 2006. 8(2): p. 110-20.

216. Regueiro, M., *et al.*, **High levels of Paleolithic Y-chromosome lineages characterize Serbia**. Gene, 2012. 498(1): p. 59-67.

217. Larmuseau, M.H., *et al.*, **Micro-geographic distribution of Y-chromosomal variation in the central-western European region Brabant**. Forensic Sci Int Genet, 2010. 5(2): p. 95-9.

218. Myres, N.M., *et al.*, **A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe**. Eur J Hum Genet, 2011.

219. Sims, L.M., D. Garvey, and J. Ballantyne, **Sub-populations within the major European and African derived haplogroups R1b3 and E3a are differentiated by previously phylogenetically undefined Y-SNPs**. Human Mutation, 2007. 28(1): p. 97-97.

220. Rootsi, S., *et al.*, **Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in europe**. Am J Hum Genet, 2004. 75(1): p. 128-37.

221. Adams, S.M., *et al.*, **The case of the unreliable SNP: recurrent back-mutation of Y-chromosomal marker P25 through gene conversion**. Forensic Sci Int, 2006. 159(1): p. 14-20.

222. Faith, S.A., *et al.*, **The application of next generation sequencing to STR typing and investiagtive genetics.** Promega Corp.

223. Irwin, J.A., *et al.*, **mtGenome reference population databases and the future of forensic mtDNA analysis**. Forensic Sci Int Genet, 2011. 5(3): p. 222-5.

224. Van Neste, C., *et al.*, **Forensic STR analysis using massive parallel sequencing**. Forensic Sci Int Genet, 2012. 6(6): p. 810-8.

225. Iseki, S., **DNA strand breaks in rat tissues as detected by in situ nick translation**. Exp Cell Res, 1986. 167(2): p. 311-26.

226. Johnstone, A.P. and G.T. Williams, **Role of DNA breaks and ADP-ribosyl transferase activity in eukaryotic differentiation demonstrated in human lymphocytes**. Nature, 1982. 300(5890): p. 368-70.

227. Di Bernardo, G., *et al.*, **Enzymatic repair of selected cross-linked homoduplex molecules enhances nuclear gene rescue from Pompeii and Herculaneum remains**. Nucleic Acids Res, 2002. 30(4): p. e16.

228. Pusch, C.M., I. Giddings, and M. Scholz, **Repair of degraded duplex DNA from prehistoric samples using *Escherichia coli* DNA polymerase I and T4 DNA ligase**. Nucleic Acids Res, 1998. 26(3): p. 857-9.

229. Singh, K., *et al.*, **Participation of the fingers subdomain of *Escherichia coli* DNA polymerase I in the strand displacement synthesis of DNA**. J Biol Chem, 2007. 282(14): p. 10594-604.

230. Xie, P., **Modeling translocation dynamics of strand displacement DNA synthesis by DNA polymerase I**. J Mol Model, 2012. 18(5): p. 1951-60.

231. DeCarlo, L., *et al.*, **Formation of purine-purine mispairs by *Sulfolobus solfataricus* DNA polymerase IV**. Biochemistry, 2008. 47(31): p. 8157-64.

232. Lawyer, F.C., *et al.*, **High-level expression, purification, and enzymatic characterization of full-length *Thermus aquaticus* DNA polymerase and a truncated form deficient in 5' to 3' exonuclease activity**. PCR Methods Appl, 1993. 2(4): p. 275-87.

233. Lawyer, F.C., *et al.*, **Isolation, characterization, and expression in *Escherichia coli* of the DNA polymerase gene from *Thermus aquaticus***. J Biol Chem, 1989. 264(11): p. 6427-37.

234. Barnes, W.M., **The fidelity of *Taq* polymerase catalyzing PCR is improved by an N-terminal deletion**. Gene, 1992. 112(1): p. 29-35.

235. Lundberg, K.S., *et al.*, **High-fidelity amplification using a thermostable DNA polymerase isolated from *Pyrococcus furiosus***. Gene, 1991. 108(1): p. 1-6.

236. Bebenek, K., *et al.*, **The fidelity of DNA synthesis catalyzed by derivatives of *Escherichia coli* DNA polymerase I**. J Biol Chem, 1990. 265(23): p. 13878-87.

237. Mattila, P., *et al.*, **Fidelity of DNA synthesis by the *Thermococcus litoralis* DNA polymerase an exttemely heat stable enzyme with proofreading activity**. Nucleic Acids Research, 1991. 19(18): p. 4967-4973.

238. Nagy, M., *et al.*, **Evaluation of whole-genome amplification of low-copy-number DNA in chimerism analysis after allogeneic stem cell transplantation using STR marker typing**. Electrophoresis, 2006. 27(15): p. 3028-37.

239. Sherry, S.T., *et al.*, **dbSNP: the NCBI database of genetic variation**. Nucleic Acids Res, 2001. 29(1): p. 308-11.

240. **The International HapMap Project**. Nature, 2003. 426(6968): p. 789-796.

241. Zweig, A.S., *et al.*, **UCSC genome browser tutorial**. Genomics, 2008. 92(2): p. 75-84.

242. Hirakawa, M., **HOWDY: an integrated database system for human genome research**. Nucleic Acids Res, 2002. 30(1): p. 152-7.

7.4 Supplemental material

Appendix-Table 7-1: Summary of all typed variations for male DNA samples.

sample ID	continent	country	file reference	typ of sample	haplogroup	MID504 Y/X = -/ACA		rs253929 Y/X = G/A		rs2757240 Y/X = T/G		rs253 4930 Y/X = G/A	rs253 4931 Y/X = G/C	rs594 0712 (A/C)	P29 (A/C)	P30 (G/A)	P38 (A/C)	P40 (C/T)	P41.2 =M359 (T/C)	M170 (A/C)	M173 (A/C)	M198 (C/T)	P224 (C/T)	P240 (T/C)	P244 (G/A)	M343 (C/A)	rs13304202 (G/A)	rs2040607 (C/T)	L12 CAGAC CTTTGC -/-	rs3572537 37 Y/X = T/A/AT	rs3448538 0 Y/X = TC/GT	rs3396332 9 Y/X = -/TTA
						Y-Chr.	X-Chr.	Y-Chr.	X-Chr.	Y-Chr.	X-Chr.	X-Chr.	X-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.	Y-Chr.
1114	Europe[1]	Germany	lab-book	Lab / Blood	I1 - P30	Del.	Ins.	A		G	A	C	C	A	A	C	T	T		A	C	C	T	G	C	Ins		Ins	TA	TA	TA
1118	Europe[1]	Germany	lab-book	Lab / Blood	R1b1b2-M269	Del.	Ins.	G	A	T	G	A	C	C	G	A	C	T		C	C	T	C	A	A	A	C		Del	TC	TA	TA	TA
1120	Europe[1]	Germany	lab-book	Lab / Blood	R1a1-M198	Del.	Ins.		A		G	A	C	C	A	G	A	C	T		T		C	A	C		C	Ins	Del	TC	TA	TA	TA
1180	Europe[1]	Germany	lab-book	Lab / Blood	I1 - P30	Del.			A		G	A	C	C	A	A	C	T	T	C			T	G	G		G	C	Ins	Del	TC	TA	TA	TA
1550	Europe[1]	Germany	lab-book	Lab / Blood	I1 - P30	Del.	Ins.		A		G	A	C	C	A	A	C	T	T		A	C		T	G	C	G	C	Ins	Del	TC	TA	TA	TA
1561	Europe[1]	Germany	lab-book	Lab / Blood	R1a1-M198	Del.									A	G	A	C	T		C	T		C	A	C	G	C	Ins	Del	TC	TA	TA	TA
1573	Europe[1]	Germany	lab-book	Lab / Blood	I1 - P30	Del.									A	A	C	T	T		A	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
3299	Europe[1]	Germany	lab-book	Lab / Blood	R1b1b2-M269	Del.	Ins.		A		G	A	C	C	A	G	A	C	T		C	C		C	G	C	A	C	Ins	Del	TC	TA	TA	TA
3300	Europe[1]	Germany	lab-book	Lab / Blood	R1b1b2-M269	Del.									A	G	A	C	T		C	C		C	A	C	A	C	Ins	Del	TC	TA	TA	TA
4202			lab-book	Lab																													
4401			lab-book	Lab																													
4471			lab-book	Lab																													
4504			lab-book	Lab																													
4518			lab-book	Lab																													
5572	West-Asia[1]		lab-book	Lab / Blood	J-P209	Del.																											
5573	Europe[1]	Germany	lab-book	Lab / Blood	I1 - P30	Del.									A	A	C	T	T		A	C		C	G	C	G	C	Ins	Del	TC	TA	TA	TA
5574	Europe[1]	Germany	lab-book	Lab / Blood	R1a1-M198	Del.									A	G	A	C	T		C	T		C	A	C	G	C	Ins	Del	TC	TA	TA	TA
5667	Europe[1]	Germany	lab-book	Lab / Blood	R1b1b2-M269	Del.	Ins.						C		A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
21151	Europe[1]	Germany	lab-book	Lab / Blood	I1 - P30	Del.			A		G	A	C		A	A	C	T	T		A	C		C	G	C	A	C	Ins	Del	TC	TA	TA	TA
21153	Europe[1]	Germany	lab-book	Lab / Blood	F-T-213	Del.									A	G	A	C	T		A	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
21154	Europe	Germany	lab-book	Lab / Blood	R1b1b2-M269	Del.	Ins.		A		G	A	C	C	A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
64407	Europe	Macedonia	2008.YA2	Immigration / Saliva / WGA		Del.	Ins.	G	A	T	G	A	C	C	A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
65991	Europe	Turkey	2008.YA3	Immigration / Saliva / WGA		Del.	Ins.		A		G	A	C	C	A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
66430	Africa	Angola	2008.YA11	Immigration / Saliva / WGA		Del.	Ins.	G	A	T	G	A	C	C	A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
66766	Africa	Nigeria	2008.YA19	Immigration / Saliva / WGA		Del.			A		G	A	C	C	A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
67211	Europe[1]	Macedonia	2008.YA28	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
67213	Africa	Ivory Coast	2008.YA30	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
69078	Europe	Turkey	2008.YA43	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
69462	Asia	Vietnam	2008.YA57	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
70147	Europe	Macedonia	2008.YA52	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
70149	Africa	Cameroon	2008.YA53	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
71149	Africa	Egypt	2008.YA60	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
71404	Europe	Turkey	2008.YA61	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
72685	Europe	Turkey	2008.YA68	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
73735	Africa	Ghana	2008.YA79	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
73737	Asia	Vietnam	2008.YA72	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
73739	Africa	Ghana	2008.YA73	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
75939	Europe	Turkey	2008.YA81	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
77067	Europe	Macedonia	2008.YA89	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
77069	West-Asia	Libanon	2008.YA88	Immigration / Saliva / WGA	E-P170	Del.		G		T																			Ins	Del	TC	TA	TA	TA
77072	Europe	Macedonia	2008.YA90	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
77074	Asia[1]	Vietnam	2008.YA87	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
78506	Asia	Vietnam	2008.YA93	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
78597	Europe	Macedonia	2008.YA95	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
80152	Africa	Nigeria	2008.YA106	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
80156	Europe	Macedonia	2008.YA104	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
80283	Europa[2]	Macedonia	2008.YA110	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
80285	Europe[1]	Macedonia	2008.YA111	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
80287	Africa[1]	Senegal	2008.YA112	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
82270	Africa[1]	Guinea	2008.YA115	Immigration / Saliva / WGA		Del.		G		T					A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
87548	Europe[1]	Turkey	2009.YA1	Immigration / Saliva / WGA		Del.	Ins.	G	A	T	G	A	C	C	A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
88159	Europe	Turkey	2009.YA6	Immigration / Saliva / WGA		Del.	Ins.	G	A	T	G	A	C	C	A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
88495	Africa	Ghana	2009.YA12	Immigration / Saliva / WGA	E-P170	Del.	Ins.	G	A	T	G	A	C	A	A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
88508	Africa	Uganda	2009.YA13	Immigration / Saliva / WGA		Del.	Ins.	G	A	T	G	A	C	C	A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA
88510	Asia	Vietnam	2009.YA14	Immigration / Saliva / WGA	E-P170	Del.	Ins.	G	A	T	G	A	C	C	A	G	A	C	T		C	C		T	G	C	A	C	Ins	Del	TC	TA	TA	TA

sample ID	continent	country	file reference	typ of sample	haplogroup	MID504 Y/X = -/ACA		rs253929 Y/X= G/A		rs2757249 Y/X = T/G		rs253 4930 G/A	rs253 4931 Y/X = G/C	rs504 0712 (A/C)	P29 (G/A)	P30 (A/C)	P38 (C/T)	P40 =M359 (T/C)	P41.2 =M359 (T/C)	MI70 (A/C)	MI73 (C/T)	M198 (C/T)	P224 (T/C)	P240 (G/A)	P244 (C/A)	M343 (G/A)	rs13304202 (C/T)	rs2040607	LL12 CAG4C CTTTGC / -	rs35753737 Y/X = T/A/AT	rs34485380 Y/X = T/C/GT	rs33963329 Y/X = -/TTA
						Del.	Ins.	Y- Chr.	X- Chr.	Y- Chr.	X- Chr.	Y- Chr.	X- Chr.	X- Chr.	X- Chr.	Y- Chr.	Y- Chr.	Y- Chr.	Y- Chr.	Y-Chr.	Y- Chr.	Y- Chr.	Y- Chr.	Y- Chr.	Y- Chr.	Y- Chr.	Y-Chr.	Y- Chr.	Y- Chr.	Y- Chr.	Y- Chr.	Y- Chr.	Y- Chr.
89075	Europe	Yugoslavia	2009YA27	Immigration / Saliva / WGA																												
89084	Europe	Turkey	2009YA23	Immigration / Saliva / WGA																												
89470	Asia[1]	Vietnam	2009YA31	Immigration / Saliva / WGA																												
89473	Africa	Nigeria	2009YA30	Immigration / Saliva / WGA																												
89715	Natv America[1]	Ecuador	lab-book	Pop / Blood	Q																											
89968	Natv America[1]	Ecuador	lab-book	Pop / Blood	Q																											
90192	Asia[1]	Vietnam	2009YA35	Immigration / Saliva / WGA																												
90198	Europe	Turkey	2009YA36	Immigration / Saliva / WGA																												
90684	Natv America	Ecuador	lab-book	Pop / Blood	Q																											
94445	Eurasia[3]		lab-book	Lab / Saliva	K-T M9																											
94446	Europe[3]		lab-book	Lab / Saliva	R-M207																											
94447	Europe	Germany	lab-book	Lab / Blood	R1b1b2b-U152																											
94448	Asia	China[1]	lab-book	Lab / Saliva	C- RPS4Y711																											
95210	Europe[1]	(Scotland)	lab-book	Lab / Blood	P-P240																											
95211	Europe[1]	Germany	lab-book	Lab / Saliva																												
99309	Europe[1]	Germany	lab-book	Lab / Saliva	R1b1b2b-U152																											
105425	Europe[1]	Panetleria	lab-book	Lab / Saliva	J-P209																											
110749	Europe[1]	Germany	lab-book	Lab / Saliva	R1b1b2 g* - U106																											
110776	Europe[1]	Germany	lab-book	Lab / Saliva	Rb1-M343																											
104122	Europe[1]	Finland	lab-book	Lab / Blood	N1C-Tat																											

(grey) characteristics of male samples, (green) all variation in the near of MID504, (orange) all SNPs, (blue) all polynucleotide polymorphisms
(1) ancestry could be confirmed by Y-chromosomal haplotype analysis, (2) Y-chromosomal haplotype showed only admixed haplotype,
(3) detailed ancestry is unknown, (Del) deletion, (Ins) Insertion

Appendix-Table 7-2: Summary of all typed variations for female DNA samples.

sample ID	typ of sample	MID504 Y/X = -		rs253929 Y/X = G/A		rs2757249 Y/X = T/G		rs2534930 Y/X = G/A	rs2534931 Y/X = G/C	rs5940712
		Y-Chr.	X-Chr.	Y-Chr.	X-Chr.	Y-Chr.	X-Chr.	X-Chr.	X-Chr.	X-Chr.
2	Lab / Blood		Ins.		A/A		G/G	A/A	C/C	A/C
767	Lab / Blood		Ins.		A/A		G/G			A/C
1117	Lab / Blood		Ins.		A/A		G/G	A/A	C/C	C/C
1552	Lab / Blood		Ins.				G/G			C/C
1554	Lab / Blood		Ins.		A/A		G/G	A/A	C/C	A/C
1576	Lab / Blood		Ins.		A/A		G/G	A/A	C/C	C/C
1577	Lab / Blood		Ins.		A/A		G/G	A/A	C/C	A/C
1579	Lab / Blood		Ins.		A/A		G/G	A/A	C/C	A/C
1581	Lab / Blood		Ins.							C/C
21149	Lab / Blood		Ins.							A/C
21150	Lab / Blood		Ins.		A/A		G/G	A/A	C/C	C/C
21152	Lab / Blood		Ins.		A/A		G/G	A/A	C/C	C/C
21155	Lab / Blood		Ins.							C/C
21564	Lab / Blood		Ins.		A/A		G/G	A/A	C/C	A/C
72635	Lab / Blood		Ins.							A/C
76317	Lab / Blood		Ins.		A/A		G/G	A/A	C/C	A/C
88220	Lab / Saliva		Ins.		A/A		G/G	A/A	C/C	A/C
88221	Lab / Saliva		Ins.		A/A		G/G	A/A	C/C	C/C
99307	Lab / Saliva		Ins.							C/C
99308	Lab / Saliva		Ins.							A/C

All female samples have European (Germany) ancestry. (Grey) characteristics of female samples, (green) all variation in the near of MID504.

The following sequences show the detailed information for the SNP analysis by Pyrosequencing. The PCR protocol is given in the case of an aberration to the standard protocol (chapter 2.8.2)

rs33963329 (TTA/-)

ACCTAAAAATCTCTGCAGTTGCTTTTAAGGACTAAAGGTGATACCAATATTTTTAAAGCATGAAAGAACAAACAAAATCCA
TATGTAAGACACAATGCAAAAAGAAAGAAGAATAAACACTATTATCTGTGTTAGTATTATCTTAGGAGAAAATTCAATGCA
GTTACCTATTTCT(TTA/)TATTATTATTATTATTATTATTATTATTATTATTATTATTATACTTTACGTTTTAGGGTACATAT
GCACAATATGCAGGTTAGTTACAAATGTATAAATGTGCCAAGCTGGTGTGC

µl	PCR - components	in PCR
2,5	10 x PCR buffer	1 x
2,5	2,5mM dNTPs	0,25 mM
0,5	$MgCl_2$	0,5 mM
1	Primer forward	0,8 mM
1	Primer reverse	0,8 mM
0,2	Taq-Gold Polymerase	0,5 U/µl
2	DNA (blood)	approx. 10ng
add H_2O to 25 µl		

95°C	10min	
94°C	15 sec	40x
55°C	15 sec	
72°C	30 sec	
72°C	10 min	

rs34485380 (GT/TC)

ATGCTTCATTCTTCATTTTCCTCCACTAGTTCAGGAGTTATACGTTCAATTTCTATTTGTGTGTGTGTGTGTGTGTGTGTGAT
GGAGTTTCACTCTTGTTGCCCAGGCTGGAGTGCAATGGTGCGATCTCAGCTCACCGCAACCTCTGCC(GT/TC)CCAGGTTCA
AGCAATTCTCCTGTCTCAGCCTCCGAGTAGCTGGGATTACAGGCATGCACTTCCACACCTGGCTAATTTTTTTTGTATTTTTA
GTAGAGACGGGGTTTCTTCATGTTGGTCAGGCTAGTCTCGAACTCCTGACCTTATGATCTGCCTACCTTGGCCTCCCAAAAT
GCTGAGATTACAGGTGTGGGCCACTGCGACCCGCCCAATTTCTA

rs35753737 (AT/TA)

ATTTTCATTGAAGGTTTACCCTAGACAGATGAGGGCAAGTTCTCATGGAAAATTCCTAAAAGTCTGTTTCTGCTGAAATAAA
GATACAACCCCACTAGCTATTAGTCTGCCCCTGGGATTTCTGAGCTATGCAAGCCCATGAGTTTGAAG(AT/TA)GAGATTTC
CTGTGTCGATTTTTTTTTCCCTCCAGTCATCCAAGGGTATTCACTAGGTCAGCCTTTCTCTTGATTATAGAGCTTGCGAAGCA
GTAGGGTTTGCACTCATGTGACTCAGAAAGTTACTTTTACAATCTCTGTTAA

P30 (G/A)and P29 (A/C)

AGTAGATAGAGATATCTGGTGGGCTGTTTGAAAAAGACAGATGACCATATAGTTAGATAATAGATAGATGATAGACAGGT
GATAGATAAGTTGATC(G/A)ATAAATAGATGATAGATGGATAGAATAGATGGTAGATAGGTAGATAGATGACAGGTATGT

AGGTAGATGGATGATAAATAGATAAGAGTCTGTAGATCAATGGACTTATAGATAGGTAGGTAAGTAGGTACACAGAGAG
AGAAAGAAACAGAGACAAATTTTAATGAATCATCTCACGTAGGTTTCAGGCTGACAGCACTGAAATCTGCAGGGCAGGCC
AGCAGTGTAGAAATTCTGGCAGGAGCTGATGTTGCAGTGTTGAGCCTGAAGGCTGAGGGACTTATGCAGAATTTCTACAC
TGTCATCTGAATCCAGATTTCCTTCTTCCTTTTGGGTCATTCTTGTTTCCTCAAGACAGTCAACTGATTAAATGAGGCCCACC
ACATTATGGAGGGCAATCTGCTGCACTCCAATCTGCTGACTTAAATCTCAATCTCATCTAAAAATACCTTT(A/C)CAGCAAC
ATCTAGTGTTTGACCAAATATCTGGGTACTGTGG

P38 (A/C)and P40 (C/T)

AATGCAGGCTGCATCTGCTTTTTAAAAGATGCCAACCTCAGGAGCAGTTGCCGCAGAGCAGGGCCTCCCTTCTTTGTAGAG
GACAATGCAGCACCTGCAGCAGTCTGCCCTGTGGGGGCTGGGAGGGTGGCTCCCGC(A/C)TGTCATCCCAGCAC(C/T)TTG
GGAGGCCAAGGAAGGTGGATCACTTGAGCCCAGGTGTTTGAGACCAGCCTGGGCAACATGG

P41.2 or M359 (C/T)

AAAGTTCAGGACAATGCAGAGATGTCACCTTCCGTCTATGGCCTTGAAGAAACCCTGGAGAGGACCAACGAGGCTGCAGG
TCTTAAGAGAGTGTAGATATTAAAATCCTTGCAGAATCTGAACATATCTCCTGAGACAAGCATAGTGATAGGG(C/T)GGGA
TTGGTTCATAGTGTAAACCAAAAATAAATTTAAAGGCCCCCCACAACCATCTGAATGGATTCCCGCCTTTGCCAGGGCACC
CTAAAACTTAACCTGAGAGTCTGGCTCATGCCATGATGGGAAGAGGAGTTGGGCATG

P170 (G/A)

AGAAGCATCTAAAGGTACGTCTTTAAGGCAACTTTGTAGACCTAACCTTTAGCAGCTAATCAAAGCCTAGGGAAG
AAAGTAACTTTAGGAGGTCTATAAATGTTTTTTGTCCCCTCCTGTGCCTCTTTCAGACTGTCTTTAAGAACTAGCACTTTCTG
TTTTTGGAACAGTTTGTATAATGTTGAATGATGGATAGTTGTTTCCTTTGGCAAACTGAG/ACCCAAAGAATCTGAGTTTTG
TATAGTGGTTTATCTGTTTGTAAACATTTATGTTTTGAAATTTGAAACCAGTGTATTTGTGGAAATGCCTTACGTAAAAGAA
ACTGTGTAGTGATAAAGACCTGCTTGCTGCTGTCTGGAAAAGTTAGCACAGCAAAAGTGGGTTTTTTTTTTTTTAGACTAGT
TAACAGTTGTTAAGATTACCACTTTT

M173

AAGAAATGTTGAACTGAAAGTTGATGCCACTTTTCAGAAAAATGGTTGTGTTTTGTACAAATTGAAATACATTGTTTAAAA
ATAAAGCACAGTACTCACTTTAGGTTTGCCATATAAATTTACTGTAACTTCCTAGAAAATTGGAAATAAAGTAAGAAAAATT
TTCTTACAATTCAAGGGCATTTAGAAC[A/C]CTTTGTCATCTGTTAATATTCAGAAATGATAAGCCAGTGTTTTGTTTTCAGG
ATCTGGGAAAACTGCAGCATTTCTTTTACCCATACTGAGTCAGATATATACAG

M198

CTATCCTGGATAGGCACTTGGGAACTTACACTCTTTAAGCCATTCCAGTCATGATGAGGTGGAATGTATCAGTAT

ACCAATTAATATTTTTGAAAGAG[C/T]TCTTTTAGGTTAATTTAAGTACAGCAATTTCTCATGTAATGTTTAGG

GAGTTTATTCTAACCTAGGCAAACGGCATGCTATCACAAGAAAGGTTTAAAGCTTTGATAAAATGGGGGAGATTT

P224 (C/T)

TTGGAAAATAACAATAAATTTGGGGGTGAATTCAGATTCTCTGGAATCTTGTCTAATTTATTTTTCTGGTTACTCTCAGAAAT

AATTTCAGAAATGAGTGTGACATCTT(C/T)CCCTGCTGACTGAAACCACTAACTTTCCACACAAGGGTAATATATATATAGG

ACATTACACCTCGTCACCTTATTCTGTGAATGCTATTCATGTTTTAACAATTCCATTGA

µl	PCR - components	in PCR
2,5	10 x PCR buffer	1 x
2,5	2,5mM dNTPs	0,25 mM
0,5	$MgCl_2$	0,5 mM
1	Primer forward	0,8 mM
1	Primer reverse	0,8 mM
0,2	Taq-Gold Polymerase	0,5 U/µl
2	DNA (blood)	approx. 10ng
add H_2O to 25 µl		

95°C 10min		
94°C 15 sec		
58°C 30 sec		40x
72°C 30 sec		
72°C 10 min		

P240 (C/T)

AATTCTCTTGGGTATCTGTCTAGAAGTGGAACTCCTGGGTCACATGTCAGTGATCTTTCAGATCAATAACGTCT(C/T)GTTCT

TTATCTGAGCCTACCAGCCAGCCCCTCCTTCACTATTGCACAGGGTTCTCTTGCAGATATGCCTCATTCTTTATCCAGACCAA

ACGCTGGTGCTACATACTCGGTCTGAATTTTGGAGATAATCCCCCAATGCTTCCAAAG

rs2740981 A/G or P244

ATGGGGGTTGGAGATGAAAGATTTACTGTCTCCTTTTCCTGTGGGACTAGAGCCCCAGGGCCTGGAAGACCCAGCATCTA

AGGTGTGCCACGGGGGAGCTTCACCCAGATGGAAACCACCTTTCCTGGCATTCTAGCGCCCTGCCCAGCAGTGCAACAGG

ACC[A/G]GGGATGGAGCTGCAGGACAATACCAGCACTTCAGGAACCGCGGTAAGGAGTAAGAGCTCAAGGCGGGAAAG

CCCATCCTATCCCATAGGGACAAAGAGGAGCTTGATCACCTCCTGGGGACCGATCACAAATCCAACCAAGAGTGGCTGGG

GGTAGCAGGAGTGGCCTGCCAATTTGGATGAACCTCATGTCCCCCTAACAAGTAAAGTATTTCACTGGTGGAGAAAATGG

GCCGATAGAGCGGCGAGTAAGAACTTGGAGCTTCCTTCCTTGCAGGAAGCT

M343

AAACCCCGCTTTCAGGTAATGTTAGCCAGCTAAATTAGATGGCTGGAGGAGGCGGTGTCTGATTCGCACAAGGCTCAGGG

TATTGGTTTGACCAGGCATGTCTTTCACGTAGCCCGAGAGAAAACTGGCCACCCTAGCCTTTTAAATATGCAAATGCAGAG

TGCCCTCGTGTTCCA[A/C]ACACCTGGAGATATGTGGGGGTGGCTATGCTGCCAGGCACGTGTT

GGGGAAAGAGCAAGAGGACAAAGGTGGGAATC

Tat (T/C)

AAAATATATAGTAGGTCTATAGATGTACTATATATTTATTTTATGTATATAGTACGTCTGTAGGAAAAATATAGTACATCTTT
GTCATTGATATGAAATATTGCCAATTTTAATTTTTCTGGGTTTAAAAAAATTATATGGACTCTGAGTGTAGACTTGTGAATTC
A(T/C)GTTGTTTTAATTTAATATTTCAGAGCACAGCAAGAGAGAAGGGTAAATATTCAGGTGATTATTTCACACTTTTACGG
CACCTTCTCAATTATGCTTACAATGGCAATATTAACATACCCAATGCTGAAGTTCTTCTTGTCAGTGAAATTGATTGGCTCAA
AAGGATTAGGGTAAGTTGAATAGTATTTAGTTTATTATGTCAT

rs2040607 (A/G)

GCACAAGAGCTCTCCAGCAACAAGGTATCCGAGACTCAGACTTCTGCTTTGGAACAAAAATGTGCACCTCCTGAAGGTTAG
CCACTTCCCATCCAGGACACTGTAAAACCCATCTAAAGTGCAAAGACAGACCAGC(A/G)GATTGAAATAGAACAGATCATG
AAACATTGATTGATAGGTGCTTAGCTTCCACATCACCTTGCCCTGAAGAAACAACTGGTGAT

rs13304202 (A/G)

CACATATATACACACATCAATCTACACACCATGTTTTTACACCAGATATATATGCACCATTGATATACCATATATACTCCATA
TATAATACACACTATATATATACATACACATATTTTGTGTGTATATATATATTATATACACACTCCATATATAATACTATATTT
GAAGAGGTTTATCTAAGGAACATTACTCAAGAG(A/G)CTTCCTCCATAAATAAAACTAATTCAGCTAATGAAATTCTGCTCA
CTAACCTGATGGTATGGGGGCTTTGGACGGAGGTAACAAAATTTTGCATGTGACAAGGGTATAAAGATTTGTGGCCAGAT
CTCAGAATCTGGCACTTGTAGGCATATGTCTACAAATTTCTGATA

Appendix-Figure 7-1: Sequencing details of the primer design for PCR an Pyrosequencing.
(yellow) = forward primer, (bright green) = sequencing primer, (grey) = reverse primer, (red) = SNP, (dark green)= overlap

```
LOCUS       sequence Nr. 7       4814 bp

   1  ttataacttt gtaaTATAAA CAAACACATG CTTACTGaat acttttgct
  51  tccctaaatt ttcttggtct ttttgactca atgtcctctg aaaagagcta
 101  tgaaatcatt tgcaagttta aaaaagaaaa aaaatgccac agaagtatta
 151  atcatctctg catttgatgg ttcattacct ttaaaacact aggtctgtcc
 201  acactgaaat atgctacatt tctctattta gttaattctt catatttaaa
 251  aaagtattta tttcattaag aacttgtcct cttgttaagt aatTgcatta
 301  ttttttgtt tgtttttgtt ctttcttttt ttttttttt tgagacagtc
 351  ttgttctgtc ccccaggctg gaatgcagtg tttccatctt ggctcactgc
 401  aacctctgcc ttccaggCTC AAGCAGTCCT CCCACCTcag cctccgaagt
 451  agctgggact acaggtgcat gccaccgtgc ccagctaact ttttactttt
 501  catagagatg ggggtctccc tacattgcct aggctggcat cgaactcctg
 551  agctcctgca atcctcttgt ctcagcctcc caaagtgctg gaactgcagg
 601  catgagccaC TGCACCTGGA CTTCATTGTg ttttatttat ttatttttga
 651  gtttttgaga cagtctcact ctgtcaccca ggctatagtg cagtggctgg
 701  tctcgaactc ctgacgaggc gaagtgatcc tcccgcccac cttggcctcc
 751  caaagtgctg ggattacagg cttgagacac cattcgtggc cttcattgtc
 801  ttttaattgc ttgtatatgt atggctatgt ttaccagagc tttataggat
 851  ctttctgtca ttctctgtca ttttaaaaaa atccagctct tggatttact
 901  taacaattgt attctttaat tttataaatg attgatttcc gcttctatct
 951  ttgtgttcAG ACAAGAGAGT tctacctgtt tgactttatc aaattctctt
1001  tttttccttt atgcttccgt ttctgtagct ttgtcttgtc ttgatgtttt
1051  tcatcttttc tggacttgct tgacagacat ggttttcttt ggtctttttc
1101  caaggtttga taagatctgt atcctcttat gattcagttg gctaaagatt
1151  tctcaaaaga attgtctaga tatCTCCAGG TGTCTGAATC AACAgtgaat
1201  ttttttttt ttttttgag acagagtctt gctccgtcac ccaggctaga
1251  gtgcagtagc gtgatctcag ctcattgcaa cctccgcctc ccaggttcaa
1301  gcaattattc tgcctcagcc acctaagtag ctgggattat aggcacctgc
1351  caccacagca agcaattttt ttttttttaa agacagcttc tcactctgtc
1401  acccaggctg gagtgcagtg gtatgatctc ggctccctgc aacctctgtc
1451  gcccaggttc aagttattgt cctgcctcag cctcctgaat agctgggatt
1501  acaggcccct gccactgtgc ctggctactt tttgtagttt tagtagagac
1551  agggtttcac catcttggcc aggtgggtct caaactcctg atctcaggtg
1601  acctgcctgc ctcggcctcc caaagtgctg ggattgcaga catgagccac
1651  catgcctggc caaaaaaata tcatttaCAG TCTCTTTTAT TTGATGGCAT
1701  tacaatagat ttttttttct tagtctctct tcctttttctt tctttttttt
1751  taattttgag acagAGTCTC ATTCTGTTGC CCAGGctgga gtgcagcagc
1801  tcaatcttgg ctcactgcaa cctctgcctc ccctgttcaa gagattctcc
1851  tgcctcagcc tcctgagtaa ctaggattac aggcgtgcat cccctcgcca
1901  ggctaatttt gtatattttt tataataagt tgaacttgaa ccacttgaat
1951  gccctgaacc tacttttttcc ttctattttt tttaatttct tttcattttt
2001  ctagaggtgg tatctcacta tgttgcatag gttggtctca aactcctggg
2051  ctcaatcaat cctcctccct cagcctccca aagtgctggg attatacaca
2101  taagccaccc cacccaacct tctttcaatt agtagcctaa tctggaattg
2151  ttttattgct acatttttgg ttttttattt acttttatgt ttatttttaat
2201  tttttaaatt tttattattt attgagatga catctcactg tgtcatctag
2251  gctggagtcc agtagcatga tctcagctca ctgcagcctc tgcctcctgg
2301  gctcaagtga tccacctgct gcgccctccc aaagtattgg gattacaggc
2351  atgaaccacc acatctggcc catttttgtT tatttatgta tttatttact
2401  tatttaatct tgaaatggag tttccctctg tcacccaggc tggagtgcag
```

Kommentar [JR1]: PCR-Primer forw. S1A

Kommentar [JR2]: Seq-Primer forw. S1B

Kommentar [JR3]: rs55885052

Kommentar [JR4]: PCR-Primer forw. S2A

Kommentar [JR5]: Seq-Primer forw. S2B

Kommentar [JR6]: Seq-Primer rev. S1B

Kommentar [JR7]: PCR-Primer rev. S1A

Kommentar [JR8]: PCR-Primer forw. S3A

Kommentar [JR9]: Sep-Primer forw. S3B

Kommentar [JR10]: Seq-Primer rev. S2B

Kommentar [JR11]: PCR-Primer rev. S2A

Kommentar [JR12]: PCR-Primer forw. S4A

Kommentar [JR13]: Seq-Primer rev. S3B and Seq-Primer forw. S4B

Kommentar [JR14]: Primer rev. S3A

Kommentar [JR15]: rs35938675

Appendix-Figure 7-2: Primer design for Sanger sequencing.
Primer pairs for the amplification and sequencing of one fragment are painted in the same colors. Additionally, the nested sequencing primers are written with lowercase letters and are underlined with a zig-zag pattern. Overlaps of primers are underlined with a line.

Appendix-Table 7-3: Probe design for tested extraction loci

probe	l	sequence	GC%	Tm NN	ΔG_M	ΔG_{MM}	$\Delta\Delta G_{M-MM}$	con_M*10^{-17}M	con_{MM}*10^{-17}M	Δcon_{M-MM}*10^{-17}M	HSE result in %	SD	n
P30 FG - DYS 439	**24**	GACAGGTGATAGATAAGTTGATCG	41	63	-11,1	-9,96	-1,15	297	285	**12**	-		
	23	ACAGGTGATAGATAAGTTGATCG	39	63	-10,7	-9,59	-1,15	295	274	**21**	-		
	22	CAGGTGATAGATAAGTTGATCG	40	61	-9,51	-8,36	-1,15	271	187	**85**	**60**		1
	21	AGGTGATAGATAAGTTGATCG	38	60	-9,05	-7,90	-1,15	247	135	**112**	**76**	1	2
	20	GGTGATAGATAAGTTGATCG	40	58	-8,16	-7,01	-1,15	165	52	**112**	**70**		1
	19	GTGATAGATAAGTTGATCG	36	55	-6,90	-5,70	-1,20	43	8	**34**	**75**	8	4
	18	TGATAGATAAGTTGATCG	33	53	-5,91	-4,76	-1,15	11	2	**9**	-		
	17	GATAGATAAGTTGATCG	35	52	-5,60	-4,41	-1,19	7	1	**6**	**63**	19	2
	16	ATAGATAAGTTGATCG	31	50	-4,92	-3,87	-1,05	3	1	**2**	-		
	15	TAGATAAGTTGATCG	33	47	-4,23	-3,18	-1,05	1	0	**1**	-		
	14	AGATAAGTTGATCG	35	49	-4,83	-3,78	-1,05	2	0	**2**	**-**		
P30 FG-1 - DYS 439	**25**	GACAGGTGATAGATAAGTTGATCGA	40	64	-11,8	-9,96	-1,81	299	285	**14**			
	24	ACAGGTGATAGATAAGTTGATCGA	38	64	-11,4	-9,59	-1,81	298	274	**24**			
	23	CAGGTGATAGATAAGTTGATCGA	39	62	-10,2	-8,36	-1,81	289	187	**102**	-		
	22	AGGTGATAGATAAGTTGATCGA	36	61	-9,71	-7,90	-1,81	278	135	**143**	-		
	21	GGTGATAGATAAGTTGATCGA	38	60	-8,81	-7,01	-1,80	230	52	**177**	**82**	3	4
	20	GTGATAGATAAGTTGATCGA	35	57	-7,51	-5,70	-1,81	93	8	**85**	-		
	19	TGATAGATAAGTTGATCGA	32	55	-6,57	-4,76	-1,81	29	2	**27**	-		
	18	GATAGATAAGTTGATCGA	33	54	-6,22	-4,41	-1,81	18	1	**17**	-		
	17	ATAGATAAGTTGATCGA	30	52	-5,57	-3,87	-1,70	7	1	**6**	-		
	16	TAGATAAGTTGATCGA	31	49	-4,79	-3,18	-1,61	2	0	**2**	-		
	15	AGATAAGTTGATCGA	33	51	-5,38	-3,78	-1,60	5	0	**5**	**-**		
P30 FA - DYS 439	**40**	TAATAGATAGATGATAGACAGGTGATAGATAAGTTGATCA	30	67	-16,1	-15,74	-0,40	300	300	**0**	**-**		
	35	GATAGATGATAGACAGGTGATAGATAAGTTGATCA	34	67	-15,2	-14,78	-0,40	300	300	**0**	-		
	30	ATGATAGACAGGTGATAGATAAGTTGATCA	33	66	-13,2	-12,82	-0,40	300	300	**0**	-		
	28	GATAGACAGGTGATAGATAAGTTGATCA	36	64	-12,1	-11,7	-0,40	299	299	**1**	-		
	26	TAGACAGGTGATAGATAAGTTGATCA	35	63	-10,6	-10,17	-0,40	294	289	**5**	-		
	25	AGACAGGTGATAGATAAGTTGATCA	36	64	-11,2	-10,76	-0,40	297	295	**2**	-		
	24	GACAGGTGATAGATAAGTTGATCA	38	62	-10,2	-9,83	-0,40	290	282	**8**	-		
	23	ACAGGTGATAGATAAGTTGATCA	35	62	-9,86	-9,46	-0,40	282	269	**13**	-		
	22	CAGGTGATAGATAAGTTGATCA	36	59	-8,63	-8,23	-0,40	214	172	**41**	**76**	5	3
	21	AGGTGATAGATAAGTTGATCA	33	58	-8,17	-7,77	-0,40	166	120	**45**	**80**	3	2
	20	GGTGATAGATAAGTTGATCA	35	56	-7,28	-6,88	-0,40	73	44	**28**	**83**	2	3
	19	GTGATAGATAAGTTGATCA	32	53	-5,97	-5,57	-0,40	13	7	**6**	**75**	6	5
	18	TGATAGATAAGTTGATCA	28	51	-5,03	-4,63	-0,40	3	2	**1**	-		
	17	GATAGATAAGTTGATCA	29	50	-4,68	-4,28	-0,40	2	1	**1**	**56**	6	3
	16	ATAGATAAGTTGATCA	25	47	-4,04	-3,74	-0,30	1	0	**0**	-		
	15	TAGATAAGTTGATCA	27	44	-3,35	-3,05	-0,30	0	0	**0**	-		
	14	AGATAAGTTGATCA	29	46	-3,95	-3,65	-0,30	1	0	**0**	-		
	13	GATAAGTTGATCA	31	43	-3,11	-2,82	-0,29	0	0	**0**	-		
	12	ATAAGTTGATCA	25	38	-2,84	-2,55	-0,29	0	0	**0**	-		
	11	TAAGTTGATCA	27	33	-2,16	-1,86	-0,30	0	0	**0**	-		
	10	AAGTTGATCA	30	36	-2,76	-2,46	-0,30	0	0	**0**	-		
	8	GTTGATCA	38	27	-1,66	-1,36	-0,30	0	0	**0**	-		
	6	TGATCA	33	7	-0,58	-0,28	-0,30	0	0	**0**	**-**		
P30 FA-1 - DYS 439	**25**	GACAGGTGATAGATAAGTTGATCAA	36	63	-10,8	-9,83	-1,01	296	282	**14**	**-**		
	24	ACAGGTGATAGATAAGTTGATCAA	33	63	-10,5	-9,46	-1,01	293	269	**24**	-		
	23	CAGGTGATAGATAAGTTGATCAA	35	60	-9,24	-8,23	-1,01	259	172	**86**	-		
	22	AGGTGATAGATAAGTTGATCAA	32	60	-8,78	-7,77	-1,01	227	120	**107**	**95**	2	2
	21	GGTGATAGATAAGTTGATCAA	33	58	-7,89	-6,88	-1,01	134	44	**89**	**84**	5	3
	20	GTGATAGATAAGTTGATCAA	30	55	-6,58	-5,57	-1,01	30	7	**23**	-		
	19	TGATAGATAAGTTGATCAA	26	52	-5,64	-4,63	-1,01	8	2	**6**	-		
	18	GATAGATAAGTTGATCAA	28	52	-5,29	-4,28	-1,01	5	1	**4**	-		
	17	ATAGATAAGTTGATCAA	24	49	-4,65	-3,74	-0,91	2	0	**1**	-		
	16	TAGATAAGTTGATCAA	25	46	-3,86	-3,05	-0,81	1	0	**0**	**-**		
P30 RC - DYS 439	**26**	TTCTATCCATCTATCATCTATTTATC	27	57	-7,43	-6,94	-0,49	86	48	**38**	**-**		
	25	TCTATCCATCTATCATCTATTTATC	28	57	-7,55	-7,06	-0,49	97	56	**42**	**86**	0	2
	24	CTATCCATCTATCATCTATTTATC	29	57	-7,63	-7,14	-0,49	106	61	**44**	92	9	2
	23	TATCCATCTATCATCTATTTATC	26	54	-6,02	-5,53	-0,49	13	7	**7**	-		
	22	ATCCATCTATCATCTATTTATC	27	55	-6,62	-6,13	-0,49	31	16	**16**	**51**	1	2
	21	TCCATCTATCATCTATTTATC	28	53	-5,47	-4,98	-0,49	6	3	**3**	-		
	20	CCATCTATCATCTATTTATC	30	54	-6,04	-5,55	-0,49	14	7	**7**	-		
	19	CATCTATCATCTATTTATC	26	49	-4,14	-3,66	-0,48	1	0	**0**	**46**	8	**2**
	18	ATCTATCATCTATTTATC	22	47	-3,56	-3,07	-0,49	0	0	**0**			

probe	l	sequence	GC%	Tm NN	ΔG_M	ΔG_{MM}	$\Delta\Delta G_{M-MM}$	$con_M * 10^{-17} M$	$con_{MM} * 10^{-17} M$	$\Delta con_{M-MM} * 10^{-17} M$	HSE result in %	SD	n
P30 RC-1 - DYS 439	**28**	ATTCTATCCATCTATCATCTATTTATCG	29	91	-9,7	-7,52	-2,18	278	94	**184**			
	27	TTCTATCCATCTATCATCTATTTATCG	30	60	-9,12	-6,94	-2,18	252	48	**204**	**-**		
	26	TCTATCCATCTATCATCTATTTATCG	31	60	-9,24	-7,06	-2,18	259	56	**203**	**70**	2	3
	25	CTATCCATCTATCATCTATTTATCG	32	60	-9,32	-7,14	-2,18	263	61	**201**	-		
	24	TATCCATCTATCATCTATTTATCG	30	57	-7,71	-5,53	-2,18	114	7	**107**	-		
	23	ATCCATCTATCATCTATTTATCG	30	59	-8,31	-6,13	-2,18	181	16	**165**	**73**	4	3
	22	TCCATCTATCATCTATTTATCG	32	56	-7,16	-4,98	-2,18	63	3	**60**	-		
	21	CCATCTATCATCTATTTATCG	33	57	-7,73	-5,55	-2,18	116	7	**109**	-		
	20	CATCTATCATCTATTTATCG	30	53	-5,83	-3,66	-2,17	10	0	**10**	-		
	19	ATCTATCATCTATTTATCG	26	52	-5,25	-3,07	-2,18	4	0	**4**	**-**		

P38 FA - DYS 437	**40**	AGCAGTCTGCCCTGTGGGGGCTGGGAGGGTGGCTCCCGCA	73	90	-37,29	-36,69	-0,60	300	300	**0**	**-**		
	35	TCTGCCCTGTGGGGGCTGGGAGGGTGGCTCCCGCA	74	89	-32,36	-31,76	-0,60	300	300	**0**	-		
	30	CCTGTGGGGGCTGGGAGGGTGGCTCCCGCA	77	86	-27,58	-26,98	-0,60	300	300	**0**	-		
	28	TGTGGGGGCTGGGAGGGTGGCTCCCGCA	75	84	-25,17	-24,57	-0,60	300	300	**0**	-		
	26	TGGGGGCTGGGAGGGTGGCTCCCGCA	77	85	-23,85	-23,25	-0,60	300	300	**0**	-		
	25	GGGGGCTGGGAGGGTGGCTCCCGCA	80	84	-23,50	-22,89	-0,61	300	300	**0**	-		
	24	GGGGCTGGGAGGGTGGCTCCCGCA	79	83	-21,92	-21,32	-0,60	300	300	**0**	-		
	23	GGGCTGGGAGGGTGGCTCCCGCA	78	81	-20,66	-20,06	-0,60	300	300	**0**	-		
	22	GGCTGGGAGGGTGGCTCCCGCA	77	80	-19,40	-18,80	-0,60	300	300	**0**	-		
	21	GCTGGGAGGGTGGCTCCCGCA	76	78	-18,14	-17,54	-0,60	300	300	**0**	**89**		**1**
	20	CTGGGAGGGTGGCTCCCGCA	75	77	-16,66	-16,06	-0,60	300	300	**0**	**87**		**1**
	19	TGGGAGGGTGGCTCCCGCA	74	74	-15,49	-14,88	-0,61	300	300	**0**	**85**	**6**	2
	18	GGGAGGGTGGCTCCCGCA	78	75	-15,42	-14,82	-0,60	300	300	**0**	-		
	17	GGAGGGTGGCTCCCGCA	76	72	-13,84	-13,24	-0,60	300	300	**0**	-		
	16	GAGGGTGGCTCCCGCA	75	70	-12,58	-12,08	-0,50	300	299	**0**	**67**		1
	15	AGGGTGGCTCCCGCA	73	71	-12,36	-11,86	-0,50	300	299	**0**	**68**	**6**	3
	14	GGGTGGCTCCCGCA	79	68	-11,49	-10,99	-0,50	298	297	**2**	-		
	13	GGTGGCTCCCGCA	77	65	-10,29	-9,79	-0,50	291	281	**10**	**54**		1
	12	GTGGCTCCCGCA	75	62	-9,13	-8,63	-0,50	252	214	**39**	**50**		1
	11	TGGCTCCCGCA	73	59	-8,29	-7,79	-0,50	179	123	**56**	**46**		1
	10	GGCTCCCGCA	80	58	-8,04	-7,54	-0,50	151	96	**55**	**50**	**5**	2
	8	CTCCCGCA	75	30	-5,18	-4,68	-0,50	4	2	**2**	-		
	6	CCCGCA	83	37	-4,65	-4,15	-0,50	2	1	**1**	**-**		

P38 FA-1 - DYS 437	**20**	TGGGAGGGTGGCTCCCGCAT	70	75	-15,93	-14,88	-1,05	300	300	**0**	**-**		
	19	GGGAGGGTGGCTCCCGCAT	73	76	-15,86	-14,82	-1,04	300	300	**0**	**86**		1
	18	GGAGGGTGGCTCCCGCAT	72	73	-14,28	-13,24	-1,04	300	300	**0**	-		
	17	GAGGGTGGCTCCCGCAT	71	71	-13,02	-12,08	-0,94	300	299	**0**	-		
	16	AGGGTGGCTCCCGCAT	69	71	-12,70	-11,86	-0,84	300	299	**1**	**79**		1
	15	GGGTGGCTCCCGCAT	73	69	-11,83	-10,99	-0,84	299	297	**2**	-		
	14	GGTGGCTCCCGCAT	71	66	-10,63	-9,79	-0,84	294	281	**14**	-		
	13	GTGGCTCCCGCAT	68	63	-9,47	-8,63	-0,84	270	214	**56**	-		
	12	TGGCTCCCGCAT	67	60	-8,63	-7,79	-0,84	214	123	**91**	**50**		1
	11	GGCTCCCGCAT	73	59	-8,38	-7,54	-0,84	189	96	**92**	**-**		

P38 FC - DYS 437	**20**	CTGGGAGGGTGGCTCCCGCC	80	78	-17,14	-15,90	-1,24	300	300	**0**	**60**		**1**
	19	TGGGAGGGTGGCTCCCGCC	78	75	-15,97	-14,73	-1,24	300	300	**0**	**60**		1
	18	GGGAGGGTGGCTCCCGCC	83	76	-15,90	-14,66	-1,24	300	300	**0**	**70**	**9**	4
	17	GGAGGGTGGCTCCCGCC	82	73	-14,32	-13,08	-1,24	300	300	**0**	-		
	16	GAGGGTGGCTCCCGCC	81	71	-13,07	-11,93	-1,14	300	299	**1**	-		
	15	AGGGTGGCTCCCGCC	80	72	-12,90	-11,71	-1,19	300	299	**1**	**72**	**3**	5
	14	GGGTGGCTCCCGCC	85	69	-11,98	-10,84	-1,14	299	296	**3**	-		
	13	GGTGGCTCCCGCC	85	67	-10,77	-9,63	-1,14	295	276	**20**	**60**		1
	12	GTGGCTCCCGCC	83	63	-9,61	-8,47	-1,14	275	198	**77**	**59**		1
	11	TGGCTCCCGCC	81	57	-7,79	-7,64	-0,15	226	107	**120**	-		
	10	GGCTCCCGCC	90	56	-7,54	-7,39	-0,15	203	82	**121**	**50**		1
	9	GCTCCCGCC	89	48	-7,05	-5,91	-1,14	55	11	**44**	-		
	8	CTCCCGCC	88	46	-5,66	-4,52	-1,14	8	1	**7**	**-**		

P38 FC-1 - DYS 437	**19**	GGGAGGGTGGCTCCCGCCT	79	78	-16,70	-14,66	-2,04	300	300	**0**	**-**		
	18	GGAGGGTGGCTCCCGCCT	78	75	-15,12	-13,08	-2,04	300	300	**0**	**49**	**4**	3
	17	GAGGGTGGCTCCCGCCT	77	73	-13,86	-11,93	-1,93	300	299	**1**	-		
	16	AGGGTGGCTCCCGCCT	75	74	-13,54	-11,71	-1,83	300	299	**1**	-		
	15	GGGTGGCTCCCGCCT	80	71	-12,67	-10,84	-1,83	300	296	**4**	-		
	14	GGTGGCTCCCGCCT	89	68	-11,46	-9,63	-1,83	298	276	**23**	-		
	13	GTGGCTCCCGCCT	77	65	-10,31	-8,47	-1,84	291	198	**93**	-		
	12	TGGCTCCCGCCT	75	63	-9,47	-7,64	-1,83	270	107	**163**	**39**		1
	11	GGCTCCCGCCT	81	62	-9,22	-7,39	-1,83	258	82	**176**	**-**		

probe	l	sequence	GC %	Tm NN	ΔG_M	ΔG_{MM}	$\Delta\Delta G_{M-MM}$	$con_M * 10^{-17} M$	$con_{MM} * 10^{-17} M$	$\Delta con_{M-MM} * 10^{-17} M$	HSE result in %	SD	n
P224FC - DYS390	40	TACTCTCAGAAATAATTTCAGAAATGAGTGTGACATCTTC	33	70	-18,71	-17,98	-0,73	300	300	0	-		
	39	ACTCTCAGAAATAATTTCAGAAATGAGTGTGACATCTTC	33	70	-19,11	-18,38	-0,73	300	300	0	-		
	35	TCAGAAATAATTTCAGAAATGAGTGTGACATCTTC	31	67	-15,56	-14,83	-0,73	300	300	0	-		
	34	CAGAAATAATTTCAGAAATGAGTGTGACATCTTC	32	68	-16,00	-15,27	-0,73	300	300	0	-		
	33	AGAAATAATTTCAGAAATGAGTGTGACATCTTC	30	67	-14,78	-14,05	-0,73	300	300	0	-		
	32	GAAATAATTTCAGAAATGAGTGTGACATCTTC	31	66	-13,88	-13,15	-0,73	300	300	0	-		
	31	AAATAATTTCAGAAATGAGTGTGACATCTTC	29	66	-13,51	-12,78	-0,73	300	300	0	-		
	30	AATAATTTCAGAAATGAGTGTGACATCTTC	30	66	-13,15	-12,42	-0,73	300	300	0	-		
	29	ATAATTTCAGAAATGAGTGTGACATCTTC	31	65	-12,78	-12,05	-0,73	300	299	0	-		
	28	TAATTTCAGAAATGAGTGTGACATCTTC	32	64	-11,89	-11,16	-0,73	299	297	2	-		
	27	AATTTCAGAAATGAGTGTGACATCTTC	33	65	-12,49	-11,76	-0,73	300	299	1	-		
	26	ATTTCAGAAATGAGTGTGACATCTTC	35	65	-12,16	-11,43	-0,73	299	298	1	-		
	25	TTTCAGAAATGAGTGTGACATCTTC	36	63	-10,97	-10,24	-0,73	297	290	7	-		
	24	TTCAGAAATGAGTGTGACATCTTC	38	63	-11,09	-10,36	-0,73	297	292	6	-		
	23	TCAGAAATGAGTGTGACATCTTC	39	63	-10,73	-9,99	-0,74	295	285	10	68	11	4
	22	CAGAAATGAGTGTGACATCTTC	41	63	-10,80	-10,07	-0,73	0	0	0	-		
	21	AGAAATGAGTGTGACATCTT**C**	38	61	-9,58	-8,85	-0,73	274	233	41	85	4	3
	20	GAAATGAGTGTGACATCTT**C**	40	59	-8,69	-7,95	-0,74	219	141	79	89	5	6
	19	AAATGAGTGTGACATCTT**C**	37	59	-8,32	-7,58	-0,74	182	100	82	91	5	10
	18	AATGAGTGTGACATCTT**C**	39	58	-7,95	-7,22	-0,73	141	68	73	89	7	7
	17	ATGAGTGTGACATCTT**C**	41	57	-7,58	-6,85	-0,73	100	43	58	88	9	8
	16	TGAGTGTGACATCTT**C**	44	54	-6,40	-5,77	-0,63	23	9	14	91	7	7
	15	GAGTGTGACATCTT**C**	47	54	-6,56	-5,93	-0,63	29	12	17	92	4	6
	14	AGTGTGACATCTT**C**	43	52	-6,02	-5,39	-0,63	13	5	8	86	6	6
	13	GTGTGACATCTT**C**	46	49	-5,15	-4,52	-0,63	4	1	2	84	6	5
	12	TGTGACATCTT**C**	42	45	-4,26	-3,63	-0,63	1	0	1	70	5	3
	11	GTGACATCTT**C**	45	44	-4,01	-3,38	-0,63	1	0	0	79	5	3
	10	TGACATCTT**C**	40	36	-2,85	-2,22	-0,63	0	0	0	57	6	3
	9	GACATCTT**C**	44	35	-2,60	-1,97	-0,63	0	0	0	54	7	3
	8	ACATCTTC	38	26	-2,06	-1,43	-0,63	0	0	0	-		
	6	CATCTTC	43	21	-0,93	-0,32	-0,61	0	0	0	-		
P224FT - DYS390	25	TTTCAGAAATGAGTGTGACATCTT**T**	32	63	-10,92	-10,22	-0,70	296	290	7	-		
	24	TTCAGAAATGAGTGTGACATCTT**T**	33	63	-11,05	-10,34	-0,71	297	291	6	-		
	23	TCAGAAATGAGTGTGACATCTT**T**	35	63	-10,68	-9,98	-0,70	295	285	10	60	9	5
	22	CAGAAATGAGTGTGACATCTT**T**	36	63	-10,76	-10,05	-0,71	295	287	9	-		
	21	AGAAATGAGTGTGACATCTT**T**	33	61	-9,53	-8,83	-0,70	272	231	41	75	5	3
	20	GAAATGAGTGTGACATCTT**T**	35	59	-8,64	-7,94	-0,70	215	139	75	84	6	5
	19	AAATGAGTGTGACATCTT**T**	31	59	-8,27	-7,57	-0,70	177	99	77	86	5	3
	18	AATGAGTGTGACATCTT**T**	33	58	-7,91	-7,20	-0,71	136	66	70	78	6	3
	17	ATGAGTGTGACATCTT**T**	35	57	-7,54	-6,84	-0,70	96	42	54	-		
	16	TGAGTGTGACATCTT**T**	38	54	-6,35	-5,75	-0,60	22	9	13	-		
P224RG - DYS390	20	GTGGTTTCAGTCAGCAGGG**G**	60	68	-13,00	-11,71	-1,29	300	299	1	78		1
	19	TGGTTTCAGTCAGCAGGG**G**	58	67	-12,03	-10,72	-1,31	299	295	4	-		
	18	GGTTTCAGTCAGCAGGG**G**	61	66	-11,68	-10,37	-1,31	299	292	7	-		
	17	GTTTCAGTCAGCAGGG**G**	58	63	-10,10	-8,79	-1,31	288	228	60	93		1
	16	TTTCAGTCAGCAGGG**G**	56	61	-9,16	-7,95	-1,21	254	141	114	-		
	15	TTCAGTCAGCAGGG**G**	60	61	-8,97	-7,76	-1,21	242	119	122	-		
	14	TCAGTCAGCAGGG**G**	64	60	-8,70	-7,49	-1,21	220	91	129	-		
	13	CAGTCAGCAGGG**G**	69	61	-8,88	-7,67	-1,23	235	110	125			
	12	AGTCAGCAGGG**G**	67	57	-7,76	-6,55	-1,25	119	28	91			
P224RA - DYS390	20	GTGGTTTCAGTCAGCAGGG**A**	55	67	-12,20	-11,70	-0,50	299	299	1	66		1
	19	TGGTTTCAGTCAGCAGGG**A**	52	65	-11,19	-10,71	-0,48	298	295	3	-		
	18	GGTTTCAGTCAGCAGGG**A**	55	65	-10,84	-10,36	-0,48	296	292	4	-		
	17	GTTTCAGTCAGCAGGG**A**	52	61	-9,26	-8,78	-0,48	260	227	33	86		1
	16	TTTCAGTCAGCAGGG**A**	50	59	-8,32	-7,94	-0,38	182	139	43	-		
	15	TTCAGTCAGCAGGG**A**	53	58	-8,13	-7,75	-0,38	161	118	43	-		
	14	TCAGTCAGCAGGG**A**	57	58	-7,86	-7,49	-0,37	130	91	39	-		

probe	l	sequence	GC%	Tm NN	ΔG_M	ΔG_{MM}	$\Delta\Delta G_{M-MM}$	$con_M * 10^{-17}M$	$con_{MM} * 10^{-17}M$	$\Delta con_{M-MM} * 10^{-17}M$	HSE result in %	SD	n
P240 FC - DYS 389II	23	ATCTTTCAGATCAATAACGTCTC	34	61	-9,72	-9,23	-0,49	279	258	20	-		
	22	TCTTTCAGATCAATAACGTCTC	36	59	-8,53	-8,05	-0,48	204	152	52	82		1
	21	CTTTCAGATCAATAACGTCTC	38	60	-9,10	-8,61	-0,49	251	212	39	-		
	20	TTTCAGATCAATAACGTCTC	35	57	-7,32	-6,83	-0,49	76	42	34	-		
	19	TTCAGATCAATAACGTCTC	36	56	-7,32	-6,83	-0,49	76	42	34	75	5	2
	18	TCAGATCAATAACGTCTC	38	56	-6,95	-6,46	-0,49	49	25	23	-		
	17	CAGATCAATAACGTCTC	41	56	-7,03	-6,54	-0,49	54	28	26	58		1
	16	AGATCAATAACGTCTC	37	52	-5,80	-5,41	-0,39	10	5	4	-		
P240 FC-1 - DYS 389II	24	ATCTTTCAGATCAATAACGTCTCG	37	65	-11,66	-9,23	-2,43	299	258	41	-		
	23	TCTTTCAGATCAATAACGTCTCG	39	62	-10,47	-8,05	-2,42	293	152	141	-		
	22	CTTTCAGATCAATAACGTCTCG	41	64	-11,04	-8,61	-2,43	297	212	85	-		
	21	TTTCAGATCAATAACGTCTCG	38	60	-9,26	-6,83	-2,43	260	42	218	82		1
	20	TTCAGATCAATAACGTCTCG	40	61	-9,26	-6,83	-2,43	260	42	218	81		1
	19	TCAGATCAATAACGTCTCG	42	60	-8,89	-6,46	-2,43	236	25	211	83	2	3
	18	CAGATCAATAACGTCTCG	44	60	-8,97	-6,54	-2,43	242	28	214	79		1
	17	AGATCAATAACGTCTCG	41	57	-7,74	-5,41	-2,33	117	5	112	-		
P240 FT - DYS 389II	23	ATCTTTCAGATCAATAACGTCTT	34	62	-9,82	-9,21	-0,61	281	257	24	-		
	22	TCTTTCAGATCAATAACGTCTT	31	59	-8,63	-8,03	-0,60	214	150	64	84	5	2
	21	CTTTCAGATCAATAACGTCTT	33	60	-9,20	-8,60	-0,60	257	211	46	-		
	20	TTTCAGATCAATAACGTCTT	30	57	-7,42	-6,82	-0,60	85	41	44	-		
	19	TTCAGATCAATAACGTCTT	32	57	-7,41	-6,81	-0,60	84	40	43	-		
	18	TCAGATCAATAACGTCTT	33	56	-7,04	-6,44	-0,60	54	25	30	-		
	17	CAGATCAATAACGTCTT	35	56	-7,12	-6,52	-0,60	60	27	33	-		
	16	AGATCAATAACGTCTT	31	52	-5,90	-5,40	-0,50	11	5	6	-		
P240 FT 1 - DYS 389II	24	ATCTTTCAGATCAATAACGTCTTG	33	63	-10,73	-9,21	-1,52	295	257	38	-		
	23	TCTTTCAGATCAATAACGTCTTG	35	61	-9,55	-8,03	-1,52	273	150	123	-		
	22	CTTTCAGATCAATAACGTCTTG	36	62	-10,12	-8,60	-1,52	288	211	77	-		
	21	TTTCAGATCAATAACGTCTTG	33	59	-8,34	-6,82	-1,52	184	41	143	86		1
	20	TTCAGATCAATAACGTCTTG	35	59	-8,33	-6,81	-1,52	183	40	143	94	5	3
	19	TCAGATCAATAACGTCTTG	37	58	-7,96	-6,44	-1,52	142	25	117	-		
	18	CAGATCAATAACGTCTTG	39	58	-8,04	-6,52	-1,52	151	27	123	-		
	17	AGATCAATAACGTCTTG	35	55	-6,82	-5,40	-1,42	41	5	36	-		
P240 RG - DYS 389II	21	GGTAGGCTCAGATAAAGAACG	47	64	-10,91	-9,77	-1,14	296	280	16	-		
	20	GTAGGCTCAGATAAAGAACG	45	61	-9,34	-8,19	-1,15	264	168	96	71	3	2
	19	TAGGCTCAGATAAAGAACG	42	59	-8,40	-7,25	-1,15	191	70	121	77	4	3
	18	AGGCTCAGATAAAGAACG	44	60	-8,87	-7,73	-1,14	234	116	118	79	1	6
	17	GGCTCAGATAAAGAACG	47	58	-7,94	-6,79	-1,15	139	39	100	74	4	2
	16	GCTCAGATAAAGAACG	43	55	-6,63	-5,59	-1,04	32	7	25	65	7	3
	15	CTCAGATAAAGAACG	40	50	-5,25	-4,20	-1,05	4	1	3	-		
	14	TCAGATAAAGAACG	35	47	-4,18	-3,13	-1,05	1	0	1	-		
	13	CAGATAAAGAACG	38	48	-4,72	-3,67	-1,05	2	0	2	-		
P240 RG-1 - DYS 389II	22	GGTAGGCTCAGATAAAGAACGA	45	64	-10,98	-9,77	-1,21	297	280	17	-		
	21	GTAGGCTCAGATAAAGAACGA	43	61	-9,40	-8,19	-1,21	267	168	99	-		
	20	TAGGCTCAGATAAAGAACGA	40	59	-8,46	-7,25	-1,21	197	70	127	73	1	2
	19	AGGCTCAGATAAAGAACGA	42	60	-8,94	-7,73	-1,21	240	116	124	73	3	2
	18	GGCTCAGATAAAGAACGA	44	58	-8,00	-6,79	-1,21	146	39	107	-		
	17	GCTCAGATAAAGAACGA	41	55	-6,70	-5,59	-1,11	35	7	28	-		
	16	CTCAGATAAAGAACGA	38	51	-5,21	-4,20	-1,01	4	1	3	-		
	15	TCAGATAAAGAACGA	33	48	-4,14	-3,13	-1,01	1	0	1	-		
	14	CAGATAAAGAACGA	36	48	-4,68	-3,67	-1,01	2	0	1	-		
P240 RA - DYS 389II	40	AGTGAAGGAGGGGCTGGCTGGTAGGC TCAGATAAAGAACA	53	80	-27,74	-27,34	-0,40	300	300	0	-		
	35	AGGAGGGGCTGGCTGGTAGGCTCAGAT AAAGAACA	54	79	-24,46	-24,06	-0,40	300	300	0	-		
	30	GGGCTGGCTGGTAGGCTCAGATAAAGA ACA	53	75	-19,57	-19,18	-0,39	300	300	0	-		
	28	GCTGGCTGGTAGGCTCAGATAAAGAAC A	50	72	-17,06	-16,66	-0,40	300	300	0	-		
	24	GCTGGTAGGCTCAGATAAAGAACA	46	67	-12,76	-12,36	-0,40	300	300	0	-		
	23	CTGGTAGGCTCAGATAAAGAACA	43	64	-11,27	-10,87	-0,40	298	296	2	-		
	22	TGGTAGGCTCAGATAAAGAACA	41	62	-10,10	-9,70	-0,40	288	278	10	-		
	21	GGTAGGCTCAGATAAAGAACA	43	62	-10,03	-9,63	-0,40	286	276	11	66	6	2
	20	GTAGGCTCAGATAAAGAACA	40	59	-8,45	-8,06	-0,39	196	153	43	82	6	2
	19	TAGGCTCAGATAAAGAACA	37	57	-7,51	-7,12	-0,39	93	60	33	87	2	6
	18	AGGCTCAGATAAAGAACA	39	58	-7,99	-7,59	-0,40	145	101	44	88	1	2
	17	GGCTCAGATAAAGAACA	41	56	-7,06	-6,66	-0,40	56	33	23	86	5	3
	16	GCTCAGATAAAGAACA	38	52	-5,75	-5,45	-0,30	9	6	3	-		
	15	CTCAGATAAAGAACA	33	48	-4,37	-4,07	-0,30	1	1	0	-		
	14	TCAGATAAAGAACA	29	44	-3,30	-3,00	-0,30	0	0	0	-		
	13	CAGATAAAGAACA	31	44	-3,84	-3,54	-0,30	1	0	0	-		
	12	AGATAAAGAACA	25	39	-2,71	-2,42	-0,29	0	0	0	-		
	11	GATAAAGAACA	27	36	-1,92	-1,62	-0,30	0	0	0	-		
	10	ATAAAGAACA	20	29	-1,65	-1,35	-0,30	0	0	0	-		
	8	AAAGAACA	25	24	-1,56	-1,27	-0,29	0	0	0	-		
	6	AGAACA	33	7	-1,07	-0,81	-0,26	0	0	0	-		

probe	l	sequence	GC%	Tm NN	ΔG_M	ΔG_{MM}	$\Delta\Delta G_{M-MM}$	$con_M * 10^{-17} M$	$con_{MM} * 10^{-17} M$	$\Delta con_{M-MM} * 10^{-17} M$	HSE result in %	SD	n
P240 RA 1 - DYS 389II	25	GCTGGTAGGCTCAGATAAAGAACAA	44	66	-12,78	-12,36	-0,42	300	300	**0**	-		
	24	CTGGTAGGCTCAGATAAAGAACAA	42	64	-11,29	-10,87	-0,42	298	296	**2**	-		
	23	TGGTAGGCTCAGATAAAGAACAA	39	62	-10,12	-9,70	-0,42	288	278	**10**	-		
	22	GGTAGGCTCAGATAAAGAACAA	41	62	-10,05	-9,63	-0,42	287	276	**11**	-		
	21	GTAGGCTCAGATAAAGAACAA	38	59	-8,47	-8,06	-0,41	198	153	**45**	-		
	20	TAGGCTCAGATAAAGAACAA	35	57	-7,53	-7,12	-0,41	95	60	**35**	-		
	19	AGGCTCAGATAAAGAACAA	37	58	-8,01	-7,59	-0,42	147	101	**46**	-		
	18	GGCTCAGATAAAGAACAA	39	56	-7,08	-6,66	-0,42	57	33	**24**	-		
	17	GCTCAGATAAAGAACAA	35	53	-5,77	-5,45	-0,32	9	6	**4**	-		
	16	CTCAGATAAAGAACAA	31	48	-4,29	-4,07	-0,22	1	1	**0**	-		
	15	TCAGATAAAGAACAA	27	45	-3,21	-3,00	-0,21	0	0	**0**	-		

P244 FG - DYS 437	40	CTGGCATTCTAGCGCCCTGCCCAGCAGTGCAACAGGACCG	65	85	-33,57	-32,51	-1,06	300	300	0	-		
	35	ATTCTAGCGCCCTGCCCAGCAGTGCAACAGGACCG	63	83	-28,68	-27,62	-1,06	300	300	0	-		
	30	AGCGCCCTGCCCAGCAGTGCAACAGGACCG	70	84	-26,75	-25,70	-1,05	300	300	0	-		
	28	CGCCCTGCCCAGCAGTGCAACAGGACCG	71	83	-24,29	-23,23	-1,06	300	300	0	-		
	26	CCCTGCCCAGCAGTGCAACAGGACCG	69	80	-21,29	-20,23	-1,06	300	300	0	-		
	25	CCTGCCCAGCAGTGCAACAGGACCG	68	78	-20,00	-18,95	-1,05	300	300	0	-		
	23	TGCCCAGCAGTGCAACAGGACCG	65	75	-17,60	-16,54	-1,06	300	300	0	49	7	2
	22	GCCCAGCAGTGCAACAGGACCG	68	76	-17,54	-16,48	-1,06	300	300	0	-		
	21	CCCAGCAGTGCAACAGGACCG	67	73	-15,73	-14,67	-1,06	300	300	0	-		
	20	CCAGCAGTGCAACAGGACCG	65	71	-14,44	-13,39	-1,05	300	300	0	-		
	19	CAGCAGTGCAACAGGACCG	63	69	-13,19	-12,13	-1,06	300	299	0	-		
	18	AGCAGTGCAACAGGACCG	61	68	-12,60	-11,54	-1,06	300	299	1	-		
	17	GCAGTGCAACAGGACCG	65	67	-11,70	-10,65	-1,05	299	294	4	-		
	16	CAGTGCAACAGGACCG	63	64	-10,17	-9,21	-0,96	289	257	32	72	4	6
	15	AGTGCAACAGGACCG	60	62	-9,65	-8,70	-0,95	276	220	56	69	3	2
	14	GTGCAACAGGACCG	64	60	-8,86	-7,91	-0,95	234	136	97	71	5	3
	13	TGCAACAGGACCG	62	58	-7,97	-7,02	-0,95	143	53	90	73	6	2
	12	GCAACAGGACCG	67	57	-7,72	-6,77	-0,95	115	38	77	69	4	2
	11	CAACAGGACCG	64	51	-6,02	-5,06	-0,96	13	3	10	50	4	3
	10	AACAGGACCG	60	48	-5,50	-4,55	-0,95	6	2	5	-		
	9	ACAGGACCG	66	46	-5,32	-4,36	-0,96	5	1	4	-		
	8	CAGGACCG	75	40	-4,18	-3,51	-0,67	1	0	1	-		
	6	GGACCG	83	28	-3,08	-3,08	0,00	0	0	0	-		

P244 FG-1 - DYS 437	19	AGCAGTGCAACAGGACCGG	63	71	-13,86	-11,54	-2,32	300	299	1	-		
	18	GCAGTGCAACAGGACCGG	66	69	-12,96	-10,65	-2,31	300	294	5	-		
	17	CAGTGCAACAGGACCGG	64	66	-11,43	-9,21	-2,22	298	257	41	-		
	16	AGTGCAACAGGACCGG	62	65	-10,81	-8,70	-2,11	296	220	75	-		
	15	GTGCAACAGGACCGG	66	63	-10,02	-7,91	-2,11	286	136	150	78		1
	14	TGCAACAGGACCGG	64	61	-9,13	-7,02	-2,11	252	53	199	75	6	2
	13	GCAACAGGACCGG	69	61	-8,88	-6,77	-2,11	235	38	197	73		1
	12	CAACAGGACCGG	67	55	-7,17	-5,06	-2,11	64	3	60	-		
	11	AACAGGACCGG	64	53	-6,66	-4,55	-2,11	33	2	32	-		

P244 FA - DYS 437	23	GCCCAGCAGTGCAACAGGACCA	61	60	-16,46	-16,41	0,05	300	300	0	57	1	2
	21	CCCAGCAGTGCAACAGGACCA	62	71	-14,54	-14,59	0,05	300	300	0	-		
	20	CCAGCAGTGCAACAGGACCA	60	69	-13,26	-13,31	0,05	300	300	0	-		
	19	CAGCAGTGCAACAGGACCA	57	67	-12,00	-12,05	0,05	299	299	0	-		
	18	AGCAGTGCAACAGGACCA	55	66	-11,41	-11,46	0,05	298	298	0	-		
	17	GCAGTGCAACAGGACCA	58	64	-10,51	-10,56	0,05	293	294	0	-		
	16	CAGTGCAACAGGACCA	56	61	-9,03	-9,08	0,05	246	249	-3	49	4	2
	15	AGTGCAACAGGACCA	53	59	-8,52	-8,57	0,05	203	208	-5	-		
	14	GTGCAACAGGACCA	57	57	-7,72	-7,77	0,05	115	120	-5	54	6	2
	13	TGCAACAGGACCA	53	54	-6,83	-6,89	0,06	42	45	-3	-		

P244 FA-1 - DYS 437	22	CCCAGCAGTGCAACAGGACCAG	63	73	-15,78	-14,54	-1,24	300	300	0	-		
	21	CCAGCAGTGCAACAGGACCAG	62	71	-14,49	-13,26	-1,23	300	300	0	-		
	20	CAGCAGTGCAACAGGACCAG	60	69	-13,23	-12,00	-1,23	300	299	1	-		
	19	AGCAGTGCAACAGGACCAG	58	68	-12,64	-11,41	-1,23	300	298	1	-		
	18	GCAGTGCAACAGGACCAG	61	66	-11,75	-10,51	-1,24	299	293	6	-		
	17	CAGTGCAACAGGACCAG	58	63	-10,22	-9,08	-1,14	290	249	40	-		
	16	AGTGCAACAGGACCAG	56	62	-9,60	-8,57	-1,03	275	208	67	73	5	3
	15	GTGCAACAGGACCAG	60	60	-8,81	-7,77	-1,04	230	120	109	76	4	3
	14	TGCAACAGGACCAG	57	58	-7,92	-6,89	-1,03	137	45	92	-		

P244 RC - DYS 437	22	ATTGTCCTGCAGCTCCATCCCC	59		-15,11	-14,36	-0,75	0	0	0	54	4	2
	18	TCCTGCAGCTCCATCCCC	66	69	-12,40	-11,64	-0,76	300	299	1	-		
	17	CCTGCAGCTCCATCCCC	70	69	-12,55	-11,80	-0,75	300	299	1	-		
	16	CTGCAGCTCCATCCCC	68	65	-10,66	-10,01	-0,65	295	286	9	-		
	15	TGCAGCTCCATCCCC	66	62	-9,62	-8,96	-0,66	275	241	34	62		1
	14	GCAGCTCCATCCCC	71	63	-9,65	-9,00	-0,65	276	244	32	68		1
	13	CAGCTCCATCCCC	69	58	-7,95	-7,29	-0,66	141	73	67	63		1
	12	AGCTCCATCCCC	67	56	-7,43	-6,78	-0,65	85	39	47	-		

probe	l	sequence	GC%	Tm NN	ΔG_M	ΔG_{MM}	$\Delta\Delta G_{M-MM}$	$con_M * 10^{-17}M$	$con_{MM} * 10^{-17}M$	$\Delta con_{M-MM} * 10^{-17}M$	HSE result in %	SD	n
P244 RC-1 - DYS 437	19	TCCTGCAGCTCCATCCCCG	68	72	-14,00	-11,64	-2,36	300	299	1	-		
	18	CCTGCAGCTCCATCCCCG	72	72	-14,15	-11,80	-2,35	300	299	1	-		
	17	CTGCAGCTCCATCCCCG	71	68	-12,26	-10,01	-2,25	300	286	14	-		
	16	TGCAGCTCCATCCCCG	69	66	-11,12	-8,96	-2,16	297	241	56	-		
	15	GCAGCTCCATCCCCG	73	66	-11,15	-9,00	-2,15	297	244	53	-		
	14	CAGCTCCATCCCCG	71	62	-9,45	-7,29	-2,16	269	73	195	71		1
	13	AGCTCCATCCCCG	69	61	-8,93	-6,78	-2,15	239	39	200	69	4	3
	12	GCTCCATCCCCG	75	58	-8,14	-5,98	-2,16	162	13	149	-		
	11	CTCCATCCCCG	72	53	-6,70	-4,55	-2,15	35	1	33	-		

P244 RT - DYS 437	22	ATTGTCCTGCAGCTCCATCCCT	55	72	-14,93	-14,24	-0,69	300	300	0	52	0	2
	21	TTGTCCTGCAGCTCCATCCCT	57	70	-13,78	-13,09	-0,69	300	300	0			
	20	TGTCCTGCAGCTCCATCCCT	60	70	-13,90	-13,21	-0,69	300	300	0			
	19	GTCCTGCAGCTCCATCCCT	63	71	-13,47	-12,79	-0,68	300	300	0			
	18	TCCTGCAGCTCCATCCCT	61	69	-12,21	-11,53	-0,68	299	299	1	-		
	17	CCTGCAGCTCCATCCCT	64	69	-12,37	-11,68	-0,69	300	299	1	-		
	16	CTGCAGCTCCATCCCT	62	65	-10,47	-9,79	-0,68	293	281	12	55	9	5
	15	TGCAGCTCCATCCCT	60	62	-9,43	-8,74	-0,69	268	224	44	-		
	14	GCAGCTCCATCCCT	64	62	-9,47	-8,78	-0,69	270	227	43	-		
	13	CAGCTCCATCCCT	61	57	-7,76	-7,07	-0,69	119	56	63	62	8	4
	12	AGCTCCATCCCT	58	55	-7,25	-6,56	-0,69	70	29	41	-		
	11	GCTCCATCCCT	63	52	-6,45	-5,77	-0,68	25	9	16	54		1
	10	CTCCATCCCT	60	45	-5,02	-4,33	-0,69	3	1	2	-		
	9	TCCATCCCT	55	39	-3,95	-3,26	0,00	1	0	0	-		

P244 RT-1 - DYS 437	19	TCCTGCAGCTCCATCCCTG	63	69	-12,79	-11,67	-1,12	300	299	1	-		
	18	CCTGCAGCTCCATCCCTG	66	70	-12,94	-11,83	-1,11	300	299	1	-		
	17	CTGCAGCTCCATCCCTG	65	65	-11,05	-9,93	-1,12	297	284	13	-		
	16	TGCAGCTCCATCCCTG	63	63	-9,90	-8,89	-1,01	283	236	47	-		
	15	GCAGCTCCATCCCTG	67	63	-9,94	-8,93	-1,01	284	239	45	-		
	14	CAGCTCCATCCCTG	57	59	-8,23	-7,22	-1,01	172	68	105	66	3	4
	13	AGCTCCATCCCTG	62	57	-7,72	-6,71	-1,01	115	35	80	-		
	12	GCTCCATCCCTG	67	54	-6,93	-5,91	-1,02	47	11	36	-		
	11	CTCCATCCCTG	64	48	-5,49	-4,48	-1,01	6	1	5	-		

probe	l	sequence	GC%	Tm NN	ΔG_M	ΔG_{MM}	$\Delta\Delta G_{M-MM}$	con_M*10^{-17}M	con_{MM}*10^{-17}M	Δcon_{M-MM}* 10^{-17}M	HSE result in %	SD	n
1304202 FG - DYS635	25	TATCTAAGGAACATTACTCAAGAGG	36	62	-10,38	-8,99	-1,39	292	243	49	-		
	24	ATCTAAGGAACATTACTCAAGAGG	38	63	-10,79	-9,40	-1,39	296	267	29	-		
	23	TCTAAGGAACATTACTCAAGAGG	39	61	-9,64	-8,25	-1,39	276	175	101	-		
	22	CTAAGGAACATTACTCAAGAGG	41	62	-10,21	-8,82	-1,39	289	230	59	-		
	21	TAAGGAACATTACTCAAGAGG	38	59	-8,59	-7,20	-1,39	210	66	144	75	9	2
	20	AAGGAACATTACTCAAGAGG	40	61	-9,19	-7,80	-1,39	256	124	132	-		
	19	AGGAACATTACTCAAGAGG	42	60	-8,86	-7,47	-1,39	234	90	144	-		
	18	GGAACATTACTCAAGAGG	44	58	-7,89	-6,50	-1,39	134	27	107	-		
	17	GAACATTACTCAAGAGG	41	55	-6,58	-5,19	-1,39	30	4	26	-		
	16	AACATTACTCAAGAGG	38	53	-6,26	-4,97	-1,29	19	3	16	-		
	15	ACATTACTCAAGAGG	40	52	-6,00	-4,71	-1,29	13	2	11	-		

1304202 FG-1 - DYS635	26	TATCTAAGGAACATTACTCAAGAGGC	39	64	-11,89	-9,12	-2,77	299	252	47	-		
	25	ATCTAAGGAACATTACTCAAGAGGC	40	65	-12,29	-9,53	-2,76	300	272	27	-		
	24	TCTAAGGAACATTACTCAAGAGGC	42	64	-11,15	-8,38	-2,77	297	189	109	-		
	23	CTAAGGAACATTACTCAAGAGGC	44	65	-11,71	-8,95	-2,76	299	240	58	-		
	22	TAAGGAACATTACTCAAGAGGC	41	62	-10,10	-7,33	-2,77	288	77	211	-		
	21	AAGGAACATTACTCAAGAGGC	43	63	-10,70	-7,93	-2,77	295	138	157	-		
	20	AGGAACATTACTCAAGAGGC	45	63	-10,37	-7,61	-2,76	292	103	188	-		
	19	GGAACATTACTCAAGAGGC	47	61	-9,40	-6,63	-2,77	267	32	235	81	11	2
	18	GAACATTACTCAAGAGGC	44	58	-8,09	-5,32	-2,77	157	5	152	-		
	17	AACATTACTCAAGAGGC	41	57	-7,77	-5,00	-2,77	120	3	118	-		
	16	ACATTACTCAAGAGGC	44	56	-7,41	-4,74	-2,67	84	2	82	-		

1304202 FA - DYS635	40	ATATTTGAAGAGGTTTATCTAAGGAACATTACTCAAGAGA	30	68	-16,87	-17,29	0,42	300	300	0	-		
	35	TGAAGAGGTTTATCTAAGGAACATTACTCAAGAGA	34	67	-15,22	-15,64	0,42	300	300	0	-		
	30	AGGTTTATCTAAGGAACATTACTCAAGAGA	33	65	-12,45	-12,86	0,41	300	300	0	-		
	28	GTTTATCTAAGGAACATTACTCAAGAGA	32	61	-10,17	-10,59	0,42	289	294	-5	-		
	26	TTATCTAAGGAACATTACTCAAGAGA	31	60	-8,94	-9,35	0,41	240	264	-25	-		
	25	TATCTAAGGAACATTACTCAAGAGA	32	59	-8,57	-8,99	0,42	208	243	-35	-		
	24	ATCTAAGGAACATTACTCAAGAGA	33	60	-8,97	-9,39	0,42	242	266	-24	-		
	23	TCTAAGGAACATTACTCAAGAGA	35	58	-7,82	-8,24	0,42	126	173	-47	-		
	22	CTAAGGAACATTACTCAAGAGA	36	59	-8,39	-8,81	0,42	190	230	-40	-		
	21	TAAGGAACATTACTCAAGAGA	33	55	-6,78	-7,20	0,42	39	66	-27	73	7	2
	20	AAGGAACATTACTCAAGAGA	35	57	-7,38	-7,80	0,42	81	124	-43	72	4	2
	19	AGGAACATTACTCAAGAGA	37	56	-7,05	-7,47	0,42	55	90	-34	77	11	4
	18	GGAACATTACTCAAGAGA	39	53	-6,07	-6,49	0,42	14	26	-12	70	13	3
	17	GAACATTACTCAAGAGA	35	49	-4,77	-5,19	0,42	2	4	-2	-		
	16	AACATTACTCAAGAGA	31	48	-4,55	-4,97	0,42	2	3	-1	-		
	15	ACATTACTCAAGAGA	33	46	-4,28	-4,70	0,42	1	2	-1	-		
	14	CATTACTCAAGAGA	36	43	-3,15	-3,57	0,42	0	0	0	-		
	13	ATTACTCAAGAGA	31	39	-2,79	-3,21	0,42	0	0	0	-		
	12	TTACTCAAGAGA	33	29	-1,92	-2,34	0,42	0	0	0	-		
	11	TACTCAAGAGA	36	33	-2,01	-2,43	0,42	0	0	0	-		
	10	ACTCAAGAGA	40	32	-2,51	-2,93	0,42	0	0	0	-		
	8	TCAAGAGA	38	16	-0,51	-0,93	0,42	0	0	0	-		

1304202 FA-1 - DYS635	26	TATCTAAGGAACATTACTCAAGAGAC	34	62	-10,54	-8,99	-1,55	294	243	50	-		
	25	ATCTAAGGAACATTACTCAAGAGAC	36	63	-10,95	-9,39	-1,56	296	266	30	-		
	24	TCTAAGGAACATTACTCAAGAGAC	38	61	-9,80	-8,24	-1,56	281	173	107	-		
	23	CTAAGGAACATTACTCAAGAGAC	39	62	-10,37	-8,81	-1,56	292	230	62	-		
	22	TAAGGAACATTACTCAAGAGAC	36	60	-8,75	-7,20	-1,55	225	66	159	76	8	2
	21	AAGGAACATTACTCAAGAGAC	38	61	-9,35	-7,80	-1,55	264	124	141	-		
	20	AGGAACATTACTCAAGAGAC	40	60	-9,02	-7,47	-1,55	245	90	156	73	10	2
	19	GGAACATTACTCAAGAGAC	42	58	-8,05	-6,49	-1,56	152	26	126	-		
	18	GAACATTACTCAAGAGAC	39	55	-6,74	-5,19	-1,55	37	4	33	-		
	17	AACATTACTCAAGAGAC	35	54	-6,42	-4,97	-1,45	24	3	21	-		
	16	ACATTACTCAAGAGAC	38	53	-6,06	-4,70	-1,36	14	2	12	-		

probe	l	sequence	GC%	Tm NN	ΔG_M	ΔG_{MM}	$\Delta\Delta G_{M-MM}$	$con_M * 10^{-17} M$	$con_{MM} * 10^{-17} M$	$\Delta con_{M-MM} * 10^{-17} M$	HSE result in %	SD	n
1304202 RC - DYS635	40	GAATTTCATTAGCTGAATTAGTTTTATTT	30	69	-17,97	-16,80	-1,17	300	300	**0**	-		
	35	TCATTAGCTGAATTAGTTTTATTTATGGA	31	68	-15,80	-14,64	-1,16	300	300	**0**	-		
	30	AGCTGAATTAGTTTTATTTATGGAGGAA	33	67	-14,08	-12,91	-1,17	300	300	**0**	-		
	28	CTGAATTAGTTTTATTTATGGAGGAAGC	32	64	-11,61	-10,44	-1,17	299	292	**6**	-		
	26	GAATTAGTTTTATTTATGGAGGAAGC	31	62	-10,37	-9,20	-1,17	292	257	**35**	-		
	25	AATTAGTTTTATTTATGGAGGAAGC	28	61	-9,73	-8,56	-1,17	279	207	**72**	-		
	24	ATTAGTTTTATTTATGGAGGAAGC	29	60	-9,36	-8,20	-1,16	265	169	**96**	-		
	23	TTAGTTTTATTTATGGAGGAAGC	30	58	-8,10	-6,94	-1,16	158	48	**110**	-		
	22	TAGTTTTATTTATGGAGGAAGC	32	59	-8,30	-7,13	-1,17	180	61	**119**	-		
	21	AGTTTTATTTATGGAGGAAGC	33	59	-8,70	-7,53	-1,17	220	95	**125**	**70**		1
	20	GTTTTATTTATGGAGGAAGC	35	58	-7,77	-6,60	-1,17	120	31	**90**	-		
	19	TTTTATTTATGGAGGAAGC	32	55	-6,78	-5,61	-1,17	39	7	**32**	**62**	15	3
	18	TTTATTTATGGAGGAAGC	33	54	-6,49	-5,32	-1,17	26	5	**21**	**61**	6	2
	17	TTATTTATGGAGGAAGC	35	53	-6,12	-4,95	-1,17	16	3	**13**	-		
	16	TATTTATGGAGGAAGC	38	52	-5,75	-4,68	-1,07	9	2	**7**	-		
	15	ATTTATGGAGGAAGC	40	53	-6,25	-5,19	-1,06	19	4	**15**	-		
	14	TTTATGGAGGAAGC	43	50	-5,21	-4,14	-1,07	4	1	**3**	-		
	13	TTATGGAGGAAGC	46	50	-5,43	-4,36	-1,07	6	1	**5**	-		
	12	TATGGAGGAAGC	50	48	-5,16	-4,10	-1,06	4	1	**3**	-		
	11	ATGGAGGAAGC	55	49	-5,67	-4,60	-1,07	8	2	**6**	-		
	10	TGGAGGAAGC	60	45	-4,62	-3,55	-1,07	2	0	**1**	-		
	8	GAGGAAGC	63	35	-3,31	-2,24	-1,07	0	0	**0**	-		
	6	GGAAGC	67	24	-2,22	-1,15	-1,07	0	0	**0**	-		

1304202 RC-1 - DYS635	40	GAATTTCATTAGCTGAATTAGTTTTATTTATGGAGGAAGCC	32	70	-19,27	-16,80	-2,47	300	300	**0**	-		
	35	TCATTAGCTGAATTAGTTTTATTTATGGAGGAAGCC	33	69	-17,10	-14,64	-2,46	300	300	**0**	-		
	30	AGCTGAATTAGTTTTATTTATGGAGGAAGCC	35	68	-15,37	-12,91	-2,46	300	300	**0**	-		
	28	CTGAATTAGTTTTATTTATGGAGGAAGCC	34	65	-12,90	-10,44	-2,46	300	292	**7**	-		
	26	AATTAGTTTTATTTATGGAGGAAGCC	31	63	-11,02	-8,56	-2,46	297	207	**90**	-		
	25	ATTAGTTTTATTTATGGAGGAAGCC	32	63	-10,66	-8,20	-2,46	295	169	**126**	-		
	24	TTAGTTTTATTTATGGAGGAAGCC	33	61	-9,40	-6,94	-2,46	267	48	**219**	**74**	17	2
	23	TAGTTTTATTTATGGAGGAAGCC	35	61	-9,59	-7,13	-2,46	274	61	**214**	-		
	22	AGTTTTATTTATGGAGGAAGCC	36	62	-9,99	-7,53	-2,46	285	95	**190**	-		
	21	GTTTTATTTATGGAGGAAGCC	38	60	-9,06	-6,60	-2,46	248	31	**217**	**81**	8	2
	20	TTTTATTTATGGAGGAAGCC	35	58	-8,07	-5,61	-2,46	154	7	**147**	**77**		1
	19	TTTATTTATGGAGGAAGCC	37	57	-7,78	-5,32	-2,46	122	5	**117**	**77**		1
	18	TTATTTATGGAGGAAGCC	39	57	-7,41	-4,95	-2,46	84	3	**81**	-		
	17	TATTTATGGAGGAAGCC	41	56	-7,04	-4,68	-2,36	54	2	**53**	-		
	16	ATTTATGGAGGAAGCC	44	57	-7,45	-5,19	-2,26	88	4	**84**	-		
	15	ATTTATGGAGGAAGCC	44	57	-7,45	-5,19	-2,26	88	4	**84**	-		
	14	TTTATGGAGGAAGCC	47	54	-6,40	-4,14	-2,26	23	1	**22**	-		
	13	TTATGGAGGAAGCC	50	54	-6,62	-4,36	-2,26	31	1	**30**	-		
	12	TATGGAGGAAGCC	54	53	-6,36	-4,10	-2,26	22	1	**21**	-		
	11	ATGGAGGAAGCC	58	54	-6,86	-4,60	-2,26	43	2	**42**	-		
	10	TGGAGGAAGCC	64	50	-5,81	-3,55	-2,26	10	0	**10**	-		
	8	GAGGAAGCC	67	43	-4,50	-2,24	-2,26	1	0	**1**	-		
	6	GGAAGCC	71	34	-3,41	-1,70	-1,71	0	0	**0**	-		

Tat FT - DYS 438	**23**	TGAGTGTAGACTTGTGAATTCAT	34	61,6	-9,95	-9,38	-0,57	285	266	**19**	**-**		
	22	GAGTGTAGACTTGTGAATTCAT	36	61,6	-9,89	-9,32	-0,57	283	263	**20**	-		
	21	AGTGTAGACTTGTGAATTCAT	33	60,6	-9,25	-8,68	-0,57	259	218	**41**	-		
	20	GTGTAGACTTGTGAATTCAT	35	58,5	-8,28	-7,71	-0,57	178	114	**64**	**77**	4	3
	19	TGTAGACTTGTGAATTCAT	31	56,2	-7,29	-6,72	-0,57	73	36	**37**	-		
	18	GTAGACTTGTGAATTCAT	33	55,5	-6,94	-6,37	-0,57	48	22	**26**	-		
	17	TAGACTTGTGAATTCAT	29	51,8	-5,68	-5,11	-0,57	8	3	**5**	-		
	16	AGACTTGTGAATTCAT	31	52,9	-6,15	-5,69	-0,46	16	8	**8**	**-**		

Tat FC - DYS 438	**22**	GAGTGTAGACTTGTGAATTCAC	41	61,7	-10,01	-9,49	-0,52	286	270	**15**	**-**		
	21	AGTGTAGACTTGTGAATTCAC	38	60,8	-9,37	-8,85	-0,52	263	230	**34**	-		
	20	GTGTAGACTTGTGAATTCAC	40	58,8	-8,40	-7,87	-0,53	191	132	**59**	**85**	7	2
	19	TGTAGACTTGTGAATTCAC	37	56,6	-7,41	-6,89	-0,52	84	45	**39**	-		
	18	GTAGACTTGTGAATTCAC	38	55,8	-7,06	-6,54	-0,52	56	28	**28**	**87**	1	3
	17	TAGACTTGTGAATTCAC	35	52,3	-5,80	-5,27	-0,53	10	4	**5**	-		
	16	AGACTTGTGAATTCAC	38	53,4	-6,27	-5,85	-0,42	19	11	**9**	-		
	15	GACTTGTGAATTCAC	40	51,2	-5,44	-5,02	-0,42	6	3	**3**	**-**		

probe	l	sequence	GC %	Tm NN	ΔG_M	ΔG_{MM}	$\Delta\Delta G_{M-MM}$	$con_M * 10^{-17}M$	$con_{MM} * 10^{-17}M$	$\Delta con_{M-MM} * 10^{-17}M$	HSE result in %	SD	n
S4: 2701 RG - DYS 437	**23**	AGGCTGTGCTATTGATGAAAATG	39	67,1	-13,01	-11,81	-1,20	300	299	**1**	**67**		1
	21	GCTGTGCTATTGATGAAAATG	41	65,5	-12,07	-10,88	-1,19	299	296	**3**	-		
	20	CTGTGCTATTGATGAAAATG	38	63,3	-10,72	-9,53	-1,19	295	272	**23**	**78**		1
	19	TGTGCTATTGATGAAAATG	35	60,4	-9,19	-8,00	-1,19	256	146	**110**	-		
	18	GTGCTATTGATGAAAATG	32	57,9	-7,98	-6,79	-1,19	144	39	**105**	-		
	17	TGCTATTGATGAAAATG	33	57,6	-7,87	-6,68	-1,19	132	34	**98**	-		
	16	GCTATTGATGAAAATG	29	54,3	-6,57	-5,38	-1,19	29	5	**24**	-		
	15	CTATTGATGAAAATG	31	53,4	-6,18	-5,09	-1,09	17	3	**14**	**-**		

S4: 2701 RA - DYS 437	**23**	AGGCTGTGCTATTGATGAAAATA	35	65,2	-11,90	-11,77	-0,13	299	299	**0**	**77**		1
	22	GGCTGTGCTATTGATGAAAATA	36	63,6	-10,97	-10,83	-0,14	296	296	**1**	-		
	21	GCTGTGCTATTGATGAAAATA	33	61,2	-9,62	-9,49	-0,13	275	270	**5**	-		
	20	CTGTGCTATTGATGAAAATA	30	58,1	-8,09	-7,96	-0,13	157	142	**15**	**91**		1
	19	TGTGCTATTGATGAAAATA	26	55,5	-6,88	-6,74	-0,14	44	37	**7**	-		
	18	GTGCTATTGATGAAAATA	28	55,1	-6,77	-6,64	-0,13	38	32	**6**	-		
	17	TGCTATTGATGAAAATA	24	51,4	-5,47	-5,34	-0,13	6	5	**1**	-		
	16	GCTATTGATGAAAATA	25	50,5	-5,08	-5,04	-0,04	3	3	**0**	**-**		

S6: 4204 FG - DYS 437	**20**	GGTCTTCCTCTGTTCCTCAG	55	66	-11,83	-11,12	-0,71	266	219	**48**	**49**	1	2
	19	GTCTTCCTCTGTTCCTCAG	53	63,5	-10,53	-9,82	-0,71	262	210	**52**	-		
	18	TCTTCCTCTGTTCCTCAG	50	61,7	-9,55	-8,84	-0,71	245	181	**64**	-		
	17	CTTCCTCTGTTCCTCAG	53	62	-9,66	-8,95	-0,71	249	186	**62**	**67**	2	4
	16	TTCCTCTGTTCCTCAG	50	57,5	-7,84	-7,23	-0,61	122	64	**58**	-		
	15	TCCTCTGTTCCTCAG	53	57,6	-7,89	-7,28	-0,61	127	67	**59**	-		
	14	CCTCTGTTCCTCAG	57	58	-8,02	-7,42	-0,60	140	78	**62**	-		
	13	CTCTGTTCCTCAG	54	52,3	-6,19	-5,58	-0,61	17	7	**10**	**-**		

S6 4204 FC - DYS 437	**20**	GGTCTTCCTCTGTTCCTCAC	55	65,9	-11,79	-10,96	-0,83	289	264	**25**	**88**	4	2
	19	GTCTTCCTCTGTTCCTCAC	53	63,4	-10,49	-9,66	-0,83	283	249	**35**	-		
	18	TCTTCCTCTGTTCCTCAC	50	61,6	-9,51	-8,68	-0,83	263	200	**63**	-		
	17	CTTCCTCTGTTCCTCAC	53	61,9	-9,62	-8,79	-0,83	267	208	**58**	**98**	1	2
	16	TTCCTCTGTTCCTCAC	50	57,4	-7,80	-7,07	-0,73	122	55	**67**	-		
	15	TCCTCTGTTCCTCAC	53	57,5	-7,85	-7,12	-0,73	127	59	**69**	-		
	14	CCTCTGTTCCTCAC	57	57,9	-7,99	-7,26	-0,73	142	69	**74**	-		
	13	CTCTGTTCCTCAC	54	52,2	-6,16	-5,43	-0,73	16	6	**11**	**-**		

Probes were designed for selected SNPs in forward (F) and reverse (R) orientation. For different probe length (l) the G/C content, melting temperature (Tm), and ΔG value were calculated using Visual Omp™. The success of separation by HSE with different probe lengths is given as the enrichment of one contributor with standard deviation and number of experiments (n). (-) Indicates no data, shades of grey represent separation success: dark grey indicates that the enrichment of one contributor is > 88% and the contributing allele appears only as a stutter, grey indicates enrichment between 70% and 88%, and light grey indicates fewer or no significant separation, < 70%. Concentrations (con) of heteroduplex of target and probes for match (M) and mismatch (MM) have been simulated for different probe lengths using Visual OMP™ with the assay parameters. Δ con presents the difference between match and mismatch concentrations.

Appendix-Table 7-4: List of Y-STR markers.

Marker	Y-Chr. position (hg18)
A7.1	19510289
DYF155_1 / MSY_1	7521567
DYF155_2	23296262
DYF383_1	24299316
DYF383_2	26481981
DYF384_1	9033829
DYF384_2	22299445
DYF385_1	24494390
DYF385_2	26286996
DYF387_1	24341084
DYF387_2	26440327
DYF391_1	16855452
DYF391_1	16972619
DYS11	12950136
DYS118A	22513038
DYS118B	23184986
DYS118C	24122846
DYS118D	26659403
DYS122	8824043
DYS12A	23991219
DYS12B	26790902
DYS132A	19252084
DYS132B	19311299
DYS139	14766043
DYS148	13117820
DYS19	10132128
DYS196	16704408
DYS197A	16849730
DYS197B	16978284
DYS198	17080456
DYS199	17605937
DYS200	17513825
DYS201	17866982
DYS203A	18242300
DYS203B	18906427
DYS204A	18132624
DYS204B	19016295
DYS205A	19115710
DYS205B	19447543
DYS206A	19257489
DYS206B	19306649
DYS207A	19252302
DYS207B	19310795
DYS208A	19176529
DYS208B	19386568
DYS209A	19171516
DYS209B	19391699
DYS210A	19217310
DYS210B	19345856
DYS211A	19128462
DYS211B	19434695
DYS212	19511655
DYS213	19803244
DYS214	20069917
DYS215	20596078
DYS217	20666316
DYS218	20980078
DYS219	21237275
DYS220	21797871
DYS221	21653451
DYS222	21586916
DYS223	21908873
DYS224	21965751
DYS225	22055241
DYS227A	22760673
DYS227B	22936937
DYS228A	9219230
DYS228B	22109521
DYS228C	22446797
DYS228D	22735263
DYS228E	22962332
DYS229A	22500016
DYS229B	23197738
DYS229C	24109808
DYS229D	26672168
DYS230 /DYS231	22387458
DYS232A	23649009
DYS232B	23882864
DYS232C	25282753
DYS232D	25498524
DYS234	2804715
DYS237A	23422171
DYS237B	25055920
DYS238A	23489371
DYS238B	25123119
DYS240A	24285133
DYS240B	26496302

Marker	Y-Chr. position (hg18)
DYS241A	24371895
DYS241B	26409494
DYS242	2751493
DYS243	16256683
DYS244A	18324886
DYS244B	18973779
DYS247A	23926008
DYS247B	26855795
DYS249	17067328
DYS250	2778511
DYS251	2906666
DYS252	2974322
DYS254	4932435
DYS255	5089513
DYS256	5674934
DYS257	8680484
DYS258	9065344
DYS260	7643503
DYS261	7482112
DYS262	7397666
DYS263	7351664
DYS265	6828304
DYS266	6866100
DYS268A	6211501
DYS268B	10339007
DYS269	10589852
DYS26A	22673726
DYS26B	23023912
DYS271	12606618
DYS272	12838346
DYS275	13518777
DYS276	14113480
DYS277	14296917
DYS278	14377065
DYS279	14790970
DYS27A	23216068
DYS27B /DYS400C	24091626
DYS27C /DYS400D	26690148
DYS28	16186988
DYS280	15053634
DYS281	16263736
DYS282	16079521
DYS288	7639942
DYS289	17330693
DYS33A	22570807
DYS33B	23126818
DYS33C	24180099
DYS33D	26601214
DYS376	2723328
DYS380	21172316
DYS382A	16850692
DYS382B	16977890
DYS385A	19261212
DYS385B	19302104
DYS386	2781435
DYS388	13257003
DYS389A	13122424
DYS390 /DYS708	15784491
DYS391	12613044
DYS392	21043399
DYS393	3191246
DYS398A	6178834
DYS398B	9239672
DYS398C	9259989
DYS398D	9280260
DYS398E	9300612
DYS398F	9919142
DYS398G	9939470
DYS398H	9959752
DYS398I	9980060
DYS399	4915644
DYS3A	6210855
DYS3B	10339813
DYS400A	22481843
DYS400B	23215719
DYS403	2631742
DYS413A /YCAIII_1	14608631
DYS413B /YCAIII_2	14676820
DYS414	9208949
DYS415	19612224
DYS425A /DYF371A	18485795
DYS425B/D YF371B	18662954

Marker	Y-Chr. position (hg18)
DYS425C /DYF371_3	24589309
DYS425D /DYF371_4	26192097
DYS434	12976599
DYS435	13006415
DYS436	13713222
DYS437 /DYS457	12977163
DYS438	13447409
DYS439	13025418
DYS440	13525393
DYS441	13491502
DYS442	13270762
DYS443 /DYS515 /DYF389	7568528
DYS444 /DYS542	17735723
DYS445	20552098
DYS446	3191741
DYS447 /DYS652	13788296
DYS448	22774681
DYS449	8278262
DYS450	8186441
DYS452	20080083
DYS453	2882843
DYS455	7026238
DYS456	4331090
DYS458	7927961
DYS459A	24488329
DYS459B	26293004
DYS461 /GATA_A7.2	19510202
DYS462	19776578
DYS463	7703660
DYS464A	23650484
DYS464B	23881563
DYS464C	25284229
DYS464D	24497223
DYS466	13146210
DYS467	13122547
DYS468	19936690
DYS469	19924501
DYS470	21179091
DYS472	15017926
DYS473	13939525
DYS474	16481132
DYS475	16421807
DYS476	16558199
DYS477	22822805
DYS478	5375063
DYS479 /DYF394	22361619
DYS480	8429474
DYS481 /DYS599	8486474
DYS482A /DYF380_1	24389690
DYS482B /DYF380_2	26391661
DYS483 /DYS426	17644303
DYS484	17558960
DYS485	20559231
DYS486A /DYF395_1	18248820
DYS486B /DYF395_1	18899977
DYS487 /DYS698	8974246
DYS488	6652708
DYS490	3503873
DYS491 /DYS602 /DYS719	15566390
DYS492 /DYS604	15923887
DYS493	7217439
DYS494	19845688
DYS495	13520787
DYS496	4615682
DYS497	16674938
DYS498A /DYF381_1	24270553
DYS498B /DYF381_2	26510842
DYS499	16163200

Marker	Y-Chr. position (hg18)
DYS500A /DYF396_1	24644335
DYS500B /DYF396_2	26137057
DYS501	16311814
DYS502	15199628
DYS503A /DYF386_1 /DYF397_1	22578078
DYS503B /DYF386_2 /DYF397_2	23119567
DYS503C /DYF386_3 /DYF397_3	24187380
DYS503D /DYF386_4 /DYF397_4	26593962
DYS504 /DYS660	2963439
DYS505	3700923
DYS506	14572787
DYS507	13121147
DYS508	16303351
DYS509	6471085
DYS510	15809462
DYS511	15814391
DYS512	15817084
DYS514	16577013
DYS603 /DYS516 /DYS709	15592479
DYS517 /DYS722	15787820
DYS518	15829511
DYS519A /DYF399_1	23506058
DYS519B /DYF399_2	25139802
DYS520 /DYS654	7790551
DYS521	10384553
DYS522	7475724
DYS523	18009647
DYS524A /DYF400_1	18409159
DYS524B /DYF400_2	18739658
DYS525	7137033
DYS526	3700765
DYS527A /DYF 388_1 /DYF401_1	24295351
DYS527B	26486154
DYS528A	18085393
DYS528B	19063442
DYS530	8453644
DYS532	8439449
DYS533	16902712
DYS534	16902515
DYS535A /DYF404_2	26417708
DYS536	2975122
DYS537	17868326
DYS539	17200529
DYS54	7857904
DYS540	17074803
DYS541	17748678
DYS543	17881896
DYS544	17091485
DYS546	17855949
DYS547	17381754
DYS548	20125189
DYS549	19979705
DYS550	19865513
DYS551	12573257
DYS552	12612457
DYS553 /DYF405	7506093
DYS554	3339892
DYS555	22253138
DYS556	21010943
DYS557	21644192
DYS558	15232892
DYS559	14206566
DYS560A /DYF407_1	14629441

Marker	Y-Chr. position (hg18)
DYS560B /DYF407_2	14656034
DYS561	15073849
DYS562	14752133
DYS563A /DYF408A	18533442
DYS563B /DYF408B	18615323
DYS563C /DYF408_3	24636709
DYS563D /DYF408_4	26144683
DYS564A /DYF409_1	19118867
DYS564B /DYF409B	19444286
DYS565	15036324
DYS566A /DYF410A /DYF390A	24850059
DYS566B /DYF390_2 /DYF411_2	25931428
DYS567	14808067
DYS568	8882682
DYS569	7334894
DYS570	6921010
DYS571	15256938
DYS572	3739839
DYS573	21473240
DYS574	15093451
DYS575	7496340
DYS576	7113492
DYS577	17434146
DYS578	20972048
DYS579	2902312
DYS58	6178024
DYS580	22871383
DYS581	12997836
DYS583	14745706
DYS584	21015115
DYS585	19827739
DYS586A /DYF411_1	19108113
DYS586B /YF411_2	19455024
DYS587	16538345
DYS588	15816702
DYS589	22895258
DYS59	6201592
DYS590	8616049
DYS591A /DYF412_1	18199631
DYS591B /DYF412_2	18949051
DYS592	7783997
DYS593	17095291
DYS594	20116396
DYS595	14926957
DYS596	8447807
DYS597 /DYS545	7775098
DYS598	8473249
DYS600 /DYS531	8526346
DYS601 /DYS648 /DYS632	14418676
DYS605 /DYS513	15951306
DYS606 /DYS640	16373922
DYS607	16923892
DYS608	17841611
DYS609	17863389
DYS610 /DYS471 /DYS657 /DYF393	22171115
DYS611	12611518
DYS612	14262146
DYS613	8802843
DYS614	8561828
DYS615	10465762
DYS616	10030220
DYS617	17591084
DYS618	21769974
DYS619	5932387
DYS620	16728397
DYS621	14162932
DYS622	16635267
DYS623	16670390

Marker	Y-Chr. position (hg18)
DYS624	16672109
DYS625	15657080
DYS626	22826604
DYS627	8710266
DYS62A	24608636
DYS62B	26173719
DYS630	19937292
DYS631	12450226
DYS633	14562037
DYS634	14737373
DYS635/C4	12889749
DYS636	21044391
DYS637	21387111
DYS638	16155069
DYS639 /DYS454	8284282
DYS641	14643826
DYS642	17481714
DYS643	15935524
DYS644	16291005
DYS649	14351988
DYS650	10463022
DYS651	7724642
DYS653 /DYS645	21737515
DYS655	57442861
DYS66	13527144
DYS67	12746823
DYS685 /DYS713	7892306
DYS688 /DYS711	8393768
DYS7	21907209
DYS703	9130290
DYS707	11253897
DYS710	17946231
DYS712	14067970
DYS714	20607288
DYS715	16171720
DYS716	11250315
DYS717	15822785
DYS718	15774364
DYS720	15706282
DYS721	15785426
DYS723	15791484
DYS724A	24561684
DYS724B	26219802
DYS725A	23660147
DYS725B	23871812
DYS725C	25293881
DYS725D	25487484
DYS726	12062503
DYS74	16205372
DYS77A	22694324
DYS77B	23003498
DYS79	26978310
DYS8	8906523
DYS80 /DYS538	17468242
DYS84	17728303
DYS86	15523396
DYS87	14813844
YAP	20071416
YCA II A	18131634
YCA II B	19017075
Y-GATA-A10	17228372
Y-GATA-A8	1330795
Y-GATA-H4	17253045

All listed Y-STR markers were used for the finding of pairs of phylogenetic SNPs and Y-STR marker which show less than 50 kb distance (Figure 4-5, Appendix-Figure 7-3). 390 Y-STRs were taken from the last recent survey from Hanson and Ballantyne 2006 of Y-STR markers from the entire Y-chromosome [215]. These Y-STR positions could be also confirmed with the family tree database. The position of DYS529 and DYS489 showed a disagreement and therefore were removed from this study. The family database contains 264 different STR markers from which 41 (green) have not been published from Hanson and Ballantyne. All markers labeled in yellow are redundantly designated loci in the publication of Hanson *et al.* 2006. All 45 markers in orange label redundantly designated markers which have been obtained from the family tree database.

haplogroup	SNP	<1kb	>1kb -25kb	>25kb - 50kb
A1	P108		DYS473; 3884	
A1a	P82			DYS265; 28478
A1b	P114		DYS275; 11272 DYS495; 9262 DYS440; 4656 DYS66; 2905	DYS441; 38547
A2	P4	DYS7; 107	DYS223; 1771	
	MEH1	DYS7; 565	DYS223; 2229	
	P248		DYS380; 21107	DYS470; 27882
	P247		DYS380; 21200	DYS470; 27975
	PK1		DYS578; 20847 DYS218; 12817 DYS556; 18046 DYS584; 22220	
	M141			DYS594; 48969 DYS548; 40176
	P3	DYS214; 56	YAP; 1443 DYS452; 10110	DYS594; 46423
	M6	DYS198; 36	DYS249; 13092 DYS549; 5617 DYS544; 11065 DYS593; 14871	
	M206		DYS276; 16642	DYS712; 28886
	M212			DYS712; 31881
	M196		DYS275; 21299 DYS495; 19289 DYS440; 14683 DYS66; 12932	DYS441; 48574
A2c	P262			DYS455; 34090
A3b	P289	DYS600; 736		DYS481; 4068 DYS614; 34746
	M190		DYS441; 13581	DYS440; 47472 DYS66; 49223 DYS495; 42866 DYS275; 40856 DYS438; 30512
	M220		DYS473; 6919	
A3b1	P291	DYS600; 642		DYS481; 40514 DYS614; 34840
A3b1a	P71	YAP; 50	DYS214; 1549 DYS452; 8617	DYS594; 44930
	P102			DYS473; 47886
A3b2	M305		DYS380; 12754 DYS470; 19529	
	M219		DYS473; 6784	
	M202		DYS275; 20109 DYS495; 18099 DYS440; 13493 DYS66; 11742	DYS441; 47384
BR	SRY10831.1		DYS376; 6152	DYS242; 34317
	M299		DYS380; 14422 DYS470; 21197	
	M139			DYS594; 49378 DYS548; 40585
B	P85		DYS275; 11195 DYS495; 9185 DYS440; 4579 DYS66; 2828	DYS441; 38470
B1	M288		DYS376; 13634	DYS242; 41799
B2a1	M218		DYS473; 6932	
B2a1a	P32		DYS581; 7416 DYS439; 20166 DYS435;1163	DYS434; 28653 DYS437; 28089
	P50		DYS581; 7414 DYS439; 20168 DYS435;1165	DYS434; 28651 DYS437; 28087
B2a2a	M43	DYS260; 232	DYS288; 3329	
	P111		DYS473; 4469	
B2b	M192		DYS275; 4879 DYS495; 2869 DYS440; 1737 DYS66; 3488	DYS441; 32154
	50f2(P)	DYS7; 754	DYS223; 2418	
B2b1	P6	DYS265; 39		DYS266; 37835
B2b4a	P70	YAP; 240	DYS214; 1739 DYS452; 8427	DYS594; 44740
B2b4b	M211		DYS712; 13962	

haplogroup	SNP	<1kb	>1kb -25kb	>25kb - 50kb
CF	P143			DYS67; 38956
CR	M294		DYS380; 17983 DYS470; 24758	
C	RPS4Y711= M130		DYS234; 9861 DYS386; 13419 DYS250; 16343	DYS242; 43361
	P255			DYS627; 34772
	P260		DYS510; 14062 DYS588; 21302 DYS512; 21684 DYS511; 18991 DYS390; 10909 DYS721; 9974 DYS517; 7580 DYS723; 3916 DYS718; 21036	DYS518; 34111 DYS717; 27385
	M216		DYS473; 7533	
C1	M8	DYS263; 130	DYS569; 16640	DYS262; 46132
	M131			DYS548; 48577
C1a	P121			DYS265; 28447
C2a	M208		DYS712; 17627	DYS276; 27883
C2a2	P54		DYS581; 3831 DYS439; 23751 DYS435; 4748 DYS437; 24504	DYS434; 25068
C3	PK2			DYS392; 35209 DYS636; 34217
	M217		DYS473; 7202	
	P44		DYS581; 7423 DYS439; 20159 DYS435;1156	DYS434; 28660 DYS437; 28096
C3d	M407		DYS234; 5693	DYS250; 31897 DYS386; 28473
C4	M347			DYS579; 35166 DYS251; 30812 DYS504; 25961 DYS252; 36844 DYS536; 37644
C4a	M210		DYS712; 17204	DYS276; 28306
C5	M356		DYS504; 15236	DYS579; 45891 DYS251; 41537 DYS252; 26119 DYS536; 26919
C6	P55		DYS581; 1124 DYS439; 26458 DYS435; 7455 DYS437; 21797	DYS434; 22361 DYS11;48824
DE	P183		DYS262; 5536	DYS263; 40468
	P167		DYS615; 4305 DYS650; 1565	DYS703; 37962
	P165		DYS609; 16921 DYS201; 13328 DYS537; 11984 DYS546; 24361	DYS608;38699
	P153		DYS543; 1586 DYS249; 3108 DYS540; 4367 DYS198; 10020 DYS544; 21048 DYS593; 24855	
	P144		DYS86; 2546	DYS491; 45540
	M203		DYS276; 12549	DYS712; 32961
D	IMS- JST021355		DYS453; 5582 DYS579; 13887 DYS251; 18241	
	M174		DYS438; 16265	DYS441; 27828
D2	P190			DYS476; 48444 DYS514; 29630 DYS622; 28624
p37.2=I2a;p37.1=D2	P37.1, P37.2		DYS581; 3856 DYS439; 23726 DYS435;4723 DYS437; 24529	DYS434; 25093
p53.2=D2a1b1; p53.1=C3e	P53.2, P53.1		DYS581; 3821 DYS439; 23761 DYS435; 4758 DYS437; 24494	DYS434; 25058
D2a1a	P42			DYS11; 41601 DYS439; 33681
			DYS434; 15138 DYS437; 14574	
			DYS581; 6099 DYS435; 14678	
D2a1b	IMS- JST022457		DYS580; 2602	DYS626; 47381
			DYS589; 21273	
D2a3	P120		DYS473; 4131	
D3a	P47	DYS435; 761	DYS581; 9340 DYS439; 18242	DYS434; 30577 DYS437; 30013

haplogroup	SNP	<1kb	>1kb -25kb	>25kb - 50kb
E	SRY4064	DYS376; 615		DYS242; 27550
	P172			DYS455; 1023
	P173			DYS455; 29285
	P168		DYS214; 8054 YAP; 6555 DYS452; 2112	DYS594; 38425
	P154	DYS523; 146		
	P156		DYS492; 5524 DYS643; 6113 DYS605; 21895	
	P155			DYS567; 39864 DYS87; 34087
	P152			DYS466; 28441
	P176			DYS601; 26242
	P174		DYS277; 21803	DYS649; 33268
	P170		DYS275; 12139 DYS495; 10129 DYS440; 5523 DYS66; 3772	DYS441; 39414
	P175			DYS442; 42709
	P29	DYS435; 800	DYS581; 9379 DYS439; 18203	DYS434; 30616 DYS437; 30052
E1	P147			DYS461; 32606 A7.1; 32519 DYS212; 31153
E1a2	P110		DYS473; 4408	
E1b1	P2	DYS214; 302	YAP; 1197 DYS452; 9864 DYS544; 19183	DYS594; 46177 DYS249; 43340
	P180		DYS593; 15377	DYS540; 35865 DYS198; 30212
	P181		DYS492; 20382	DYS643; 32019 DYS605; 47801
	DYS391p	DYS389A; 22 DYS467; 145	DYS466; 23808	
	P179		DYS271; 2949	DYS271; 36310 DYS611; 41210 DYS552; 42149 DYS391; 42736
E1b1a	P211	DYS376; 379		DYS242; 27786
	P293		DYS122; 11135	DYS613; 32335 DYS568; 47504
	M2=SY81	DYS271; 38	DYS611; 4938 DYS552; 5877 DYS391; 6464	DYS551; 33323
	P189			DYS67; 38846
	P1	YAP; 139	DYS214; 1638 DYS452; 8528	DYS594; 44841
	P182		DYS84; 17538 DYS444; 10118 DYS541; 2837	
E1b1a6	M195	DYS440; 537	DYS275; 7153 DYS495; 5143 DYS66; 1214	DYS441; 34428
E1b1a7	U247=P253		DYS289; 18076	DYS547; 32985
	U186			DYS472; 29980 DYS565; 48378
	M191=P86		DYS275; 10230 DYS495; 8220 DYS440; 3614 DYS66; 1863	DYS441; 37505
E1b1a7a	U174=P252		DYS634; 23378 DYS583; 15045 DYS562; 8618 DYS139; 2955	DYS279; 30219 DYS567; 47316
	p9.2=E1b1a7a1;p9.1=CR		DYS438; 11566	
E1b1a7a3a	P113			DYS265; 28133
E1b1a8	U175		DYS583; 17382 DYS562; 10955 DYS139; 5292	DYS634; 25715 DYS279; 27882 DYS567; 44979

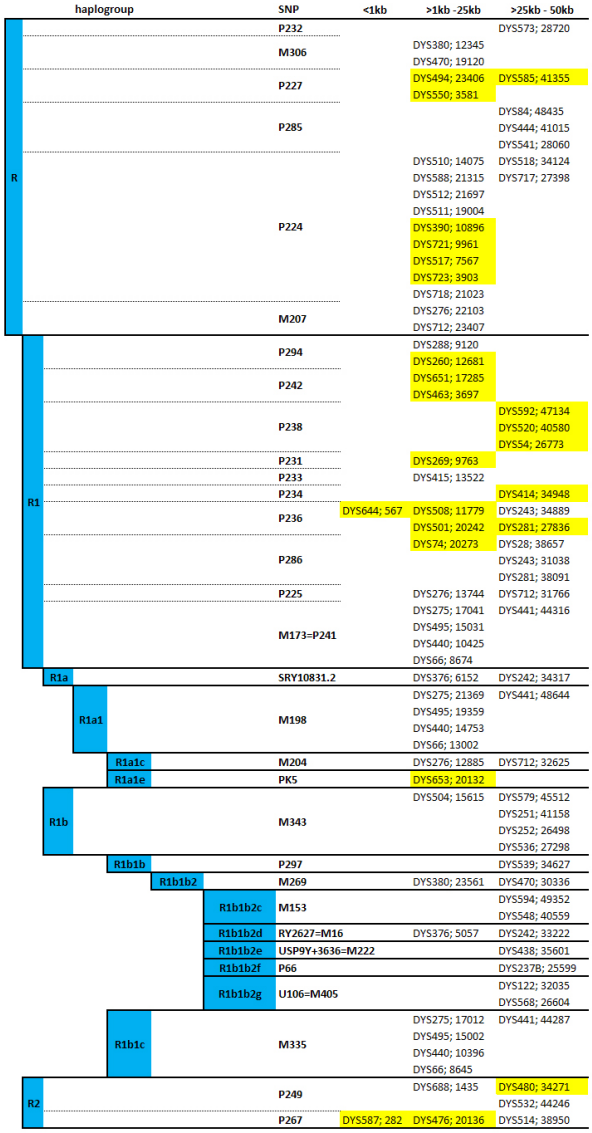
haplogroup	SNP	<1kb	>1kb -25kb	>25kb - 50kb
E1b1a8a	P278	DYS600; 707		DYS481; 40579 DYS614; 34775
	U209		DYS510; 5110 DYS588; 12350 DYS512; 12732 DYS717; 18433 DYS511; 10039 DYS390; 19861 DYS721; 18926 DYS517; 16532 DYS723; 12868	DYS718; 29988 DYS518; 25159
	P277		DYS712; 20639 DYS276; 24871	
E1b1a8a1	U290		DYS594; 10950 DYS548; 19743	DYS214; 35529 YAP; 34030 DYS452; 25363
E1b1a8a1a	U181			DYS595; 43137
E1B1a8a2	P59		DYS434; 17946 DYS437; 17982 DYS581; 3291 DYS435; 11870	DYS11; 44409 DYS439; 30873
E1b1a9	P269		DYS230; 11608	DYS479; 37447 DYS228C; 47731
	P268		DYS230; 11607	DYS479; 37446 DYS228C; 47732
E1b1b	M215			DYS473; 37693
E1b1b1a1	V12		DYS266; 16999	DYS570; 37911
E1b1b1a1b	V32			DYS455; 33417
E1b1b1a2	V36		DYS266; 8146	DYS265; 45942 DYS570; 46764
	V13		DYS570; 18747	DYS266; 36163
E1b1b1a2a	V27			DYS570; 35041
E1b1b1a2b	P65			DYS237B; 25658
E1b1b1a3	V22		DYS570; 1053	
E1b1b1a4	V65		DYS510; 12396 DYS588; 19636 DYS512; 20018 DYS511; 17325 DYS390; 12575 DYS721; 11640 DYS517; 9246 DYS723; 5582 DYS718; 22702	DYS518; 32445 DYS717; 25719
E1b1b1b2	M183			DYS438; 49232
E1b1b1e	V6			DYS455; 34231
E1b1b1f	P72	DYS214; 324	YAP; 1175 DYS452; 9842	DYS594; 46155
E1b1c	M329			DYS579; 33215 DYS251; 28861 DYS504; 27912 DYS252; 38795 DYS536; 39595
E2	P68		DYS438; 11660	
E2a	M41=P210	DYS376; 561		DYS242; 27604
E2b1a	M200		DYS275; 22033 DYS495; 20023 DYS440; 15417 DYS66; 13666	DYS441; 49308
E2b1a2	P258		DYS275; 21775 DYS495; 19765 DYS440; 15159 DYS66; 13408	DYS441; 49050

haplogroup	SNP	<1kb	>1kb -25kb	>25kb - 50kb
F-R	P141			DYS455; 25020
	P134		DYS522; 19918	DYS261; 26306
				DYS575; 40534
	P145		DYS598; 10840	DYS532; 44640
			DYS481; 2385	DYS596; 36282
				DYS530; 30445
				DYS600; 42257
	P160		DYS600; 7843	DYS481; 47715
				DYS614; 27639
	P146		DYS257; 18069	DYS590; 46366
				DYS627; 47851
	P151			DYS627; 30395
	P187			DYS414; 40697
	P138			DYS67; 37539
	P157		DYS227A; 8646	DYS228D; 34056
			DYS448; 5362	
	P136		DYS573; 23205	
	P135		DYS214; 8327	DYS594; 38152
			YAP; 6828	
			DYS452; 1839	
			DYS608; 17398	
	P148		DYS609; 4380	
			DYS201; 7973	
			DYS537; 9317	
			DYS546; 3060	
			DYS543; 22887	
			DYS198; 7414	
	P149		DYS249; 20542	
			DYS540; 13067	
			DYS544; 3615	
			DYS593; 7421	
			DYS623; 17095	DYS620; 40912
	P161		DYS624; 15376	
			DYS497; 12547	
			DYS196; 16923	
				DYS476; 48446
	P159			DYS514; 29632
				DYS622; 28622
	P14		DYS492; 15895	DYS643; 27532
				DYS605; 43314
			DYS510; 11907	DYS718; 47005
			DYS588; 4667	DYS390; 36878
	P140		DYS512; 4285	DYS721; 35943
			DYS717; 1416	DYS517; 33549
			DYS518; 8142	DYS723; 29885
			DYS511; 6978	
			DYS390; 19079	DYS510; 44050
	P166		DYS721; 20014	DYS723; 26072
			DYS517; 22408	DYS511; 48979
			DYS718; 8952	
		DYS562; 423	DYS634; 14337	DYS279; 39260
	P163		DYS583; 6004	
			DYS139; 14333	
	P133		DYS615; 10978	
			DYS650; 13718	
	M213=P137			DYS712; 31825
F1	P104			DYS473; 47694
F2	M427		DYS617; 10907	DYS484; 43031
			DYS199; 3946	DYS483; 42312
	M428		DYS617; 10869	DYS484; 42993
			DYS199; 3984	DYS483; 42350
G	P257		DYS11; 7200	DYS434; 33663
				DYS437; 34227
G1	M342	DYS221; 121	DYS557; 9138	
	M285		DYS380; 21188	DYS470; 27963
G2	P287		DYS445; 20613	DYS485; 27746
G2a	P15	DYS221; 37	DYS557; 9222	
		DYS211B; 117	DYS205B; 12965	DYS208B; 48010
G2a1	P16		DYS564B; 9708	
			DYS586B; 20446	
	M286		DYS380; 21129	DYS470; 27904
G2b	M287		DYS380; 21158	DYS470; 27933
G2c	M377		DYS275; 18050	DYS441; 45325
			DYS495; 16040	
			DYS440; 11434	
			DYS66; 9683	
H1a1	M197		DYS275; 16215	DYS441; 43490
			DYS495; 14205	
			DYS440; 9599	
			DYS66; 7848	
H1a3	M138			DYS548; 47054
H1b	M370		DYS504; 14841	DYS579; 46286
				DYS251; 41932
				DYS252; 25724
				DYS536; 26524
H2	Apt		DYS7; 1490	
			DYS223; 3154	
H2a	P80			DYS265; 28447
H2b	P266			DYS265; 28566
IJ	P130		DYS590; 2920	
	P127			DYS590; 34703
				DYS257; 29732
	P125=M429		DYS551; 31923	
				DYS271; 47975
	P129			DYS611; 43075
				DYS552; 42136
				DYS391; 41549
	P123			DYS483; 31952
	P124		DYS484; 11264	DYS200; 33871
				DYS617; 43388
I	M258		DYS275; 13981	DYS441; 41256
			DYS495; 11971	
			DYS440; 7365	
			DYS66; 5614	
	P38		DYS434; 17788	DYS11; 44251
			DYS437; 17224	DYS439; 31031
			DYS581; 3449	
			DYS435; 12028	

haplogroup	SNP	<1kb	>1kb -25kb	>25kb - 50kb
I1	M450			DYS443; 40387
				DYS288; 31027
				DYS260; 34588
	P40		DYS434; 17809	DYS11; 44266
			DYS437; 17239	DYS439; 31016
			DYS581; 3434	
			DYS435; 12013	
	M253		DYS275; 13324	DYS441; 40599
			DYS495; 11314	
			DYS440; 6708	
			DYS66; 4957	
	P30	DYS435; 346	DYS581; 8925	DYS434; 30162
			DYS439; 18657	DYS437; 29598
I1b	M227		DYS276; 12640	DYS712; 32870
I1c	P109		DYS473; 4126	
I1d	P259		DYS276; 12612	DYS712; 32898
P41.2=I2A1; P41.1=D2	P41.1=P41.2		DYS581; 3843	DYS434; 25080
			DYS439; 23739	
			DYS435; 4736	
			DYS437; 24516	
I2b	P220		DYS580; 13674	
			DYS589; 10201	
	U250=P222		DYS547; 15840	DYS577; 36552
	P214=M436		Y-GATA-H4; 3842	Y-GATA-A10; 2851
	P223		DYS502; 9100	DYS571; 48210
			DYS558; 24164	
	P219			DYS712; 40725
	P217		DYS463; 15176	DYS288; 48542
				DYS651; 36158
				DYS260; 44981
			DYS688; 19939	
	P221		DYS480; 15767	DYS532; 25742
				DYS596; 34100
				DYS530; 39937
I2b1	M284		DYS380; 12471	
			DYS470; 19246	
m379= Ib2b; m201= G	M379=M201		DYS275; 18146	DYS441; 45421
			DYS495; 16136	
			DYS440; 11530	
			DYS66; 9779	
I2b3	P78			DYS265; 27917
J	M304		DYS380; 13075	
			DYS470; 19850	
	P209			DYS483; 44426
				DYS84; 39574
				DYS444; 46994
J1	M267		DYS380; 21110	DYS470; 27885
J1a	M62		DYS288; 3312	
		DYS260; 249		
M365 = J1b M390=J1c M	M365 = J1b M390=J1c M		DYS504/DYS660;	DYS579; 46366
				DYS251; 42012
				DYS252; 25644
				DYS536; 26444
J1d	P56	DYS581; 490	DYS434; 20747	DYS11; 47210
			DYS437; 20183	DYS439; 28072
			DYS435; 9069	
J1e	P58		DYS434; 20076	DYS11; 46539
			DYS437; 19512	DYS439; 28743
			DYS581; 1161	
			DYS435; 9740	
J1e1	M367		DYS504; 14811	DYS579; 46316
				DYS251; 41962
				DYS252; 25694
				DYS536; 26494
	M368		DYS504; 14807	DYS579; 46320
				DYS251; 41966
				DYS252; 25690
				DYS536; 26490
	M369		DYS504; 14962	DYS579; 46165
				DYS251; 41811
				DYS252; 25845
				DYS536; 26645
J2	M172		DYS441; 12474	DYS440; 46365
				DYS66; 48166
				DYS495; 41759
				DYS275; 39749
				DYS438; 31619
J2a	M410		DYS234; 6963	DYS386; 30243
				DYS250; 33167
J2a1	M322			DYS473; 39609
J2a7	M318		DYS380; 17935	
			DYS470; 24710	
J2a8	M319			DYS473; 37654
J2a9	M339		DYS504; 22072	DYS579; 39055
				DYS251; 34701
				DYS252; 32955
				DYS536; 33755
J2a11	M419			DYS473; 37775
J2a12	P81			DYS265;28448
J2a13	P279		DYS504; 22062	DYS579; 39065
				DYS251; 34711
				DYS252; 32945
				DYS536; 33745
J2b	M12		DYS288; 3538	
		DYS260; 23		
	M314		DYS380; 9849	
			DYS470; 16624	
	M221		DYS473; 6185	
J2b1	M205		DYS276; 16577	DYS712; 28933
J2b2	M241	DYS66; 709	DYS275; 9076	DYS441; 36351
			DYS495; 7066	
			DYS440; 2460	
J2b2c	M321		DYS275; 21495	DYS441; 48770
			DYS495; 19485	
			DYS440; 14879	
			DYS66; 13128	
J2b2d	P84			DYS265; 28537

haplogroup	SNP	<1kb	>1kb -25kb	>25kb - 50kb
K-R	P132			DYS267; 29577
	P128		DYS385B; 5163 DYS206B; 9708 DYS207B; 13854 DYS132B; 14358	DYS132A; 44857 DYS207A; 44639 DYS206A; 39452 DYS210B; 48915
	P131			DYS473; 42732
K2	P60		DYS434; 17880 DYS437; 17316 DYS581; 3357 DYS435; 11936	DYS11; 44343 DYS439; 30939
K3	P79			DYS265; 27927
K4	P263		DYS415; 13133	
L	M295		DYS380; 17877 DYS470; 24652	
	M185			DYS438; 33156
	M22			DYS388; 42554 DYS442; 28795
L2	M317		DYS380; 8257 DYS470; 15032	
L2a	M349		DYS504/DYS660; 21045 DYS252; 21045 DYS536; 21845	DYS251; 46611
L2b	M274			DYS380; 25127 DYS470; 31902
L3	M357		DYS504; 15187	DYS579; 45940 DYS251; 41586 DYS252; 26070 DYS536; 26870
L3a	PK3			DYS392; 35514 DYS636; 34522
M	P256			DYS627; 34964
M1	M4	DYS234; 87	DYS386; 23193 DYS380; 17807 DYS470; 24582	DYS250; 26117
	M296		DYS438; 7690 DYS438; 14554	DYS441; 36403
	M189	DYS214; 583	YAP; 2082 DYS452; 10749	DYS594; 47062
M1a1	P51		DYS581; 4214 DYS439; 23368 DYS435; 4365 DYS437; 24887	DYS434; 25451
M1b	P87		DYS275; 10336 DYS495; 8326 DYS440; 3720 DYS66; 1969	DYS441; 37611
M1b1	P22=M104	DYS257; 892		DYS627; 30674
M1b1a	M16	DYS214; 228	YAP; 1727 DYS452; 10394	DYS594; 46707
M2	M387		DYS234; 4441 DYS386; 18839 DYS250; 21763 DYS510; 9199	DYS242; 48781 DYS718; 44297
	M353		DYS588; 1959 DYS512; 1577 DYS717; 4124 DYS518; 10850 DYS511; 4270	DYS390; 34170 DYS721; 33235 DYS517; 30841 DYS723; 27177
M3	P118		DYS275; 10423 DYS495; 8413 DYS440; 3807 DYS66; 2056	DYS441; 37698
NO	P188		DYS225; 11491	
	P195			DYS392; 31251 DYS636; 30259
	P193		DYS217; 7293	
	P194		DYS634; 24999	DYS413B; 35554 DYS583; 33332 DYS562; 39759
	M214			DYS473; 41794
N	M231			DYS473; 39593
N1c	P105 Tat		DYS473; 7169 DYS438; 15432	
N1c1b	P67			DYS237B; 25562
N1c1c	P119		DYS438; 11759	
O	P186		DYS288; 11374 DYS260; 14935	
	M175			DYS712; 49870
	P191		DYS473; 15016	
	P196		DYS612; 1561	DYS277; 33210
O1a1	P203		DYS380; 11977 DYS470; 18752	
O2	P31		DYS581; 7415 DYS439; 20167 DYS435;1164	DYS434; 28652 DYS437; 28088
O2a1a	PK4			DYS462; 31228
O2b	SRY465		DYS376; 8148	DYS242; 36313
	IMS-JST022454		DYS589; 24711 DYS227B; 16968	DYS580; 48586 DYS228E; 42363
	P49	DYS435; 0	DYS581; 8579 DYS439; 19003	DYS434; 29816 DYS437; 29252
O2b1	47z		DYS490; 7431	
O3	P198		DYS491; 3225	DYS516; 29314
O3a	M324		DYS453; 1057 DYS579; 20526 DYS251; 24880	
	P197			DYS483; 44554 DYS84; 39446 DYS444; 46866
	P199			DYS539; 44093
	P200		DYS249; 2071 DYS540; 5404 DYS198; 11057 DYS544; 22086	DYS593; 25892

haplogroup	SNP	<1kb	>1kb -25kb	>25kb - 50kb
O3a3	P201=IMSJST021354		DYS453; 5353 DYS579; 14116 DYS251; 18470	
O3a3b1	M209		DYS712; 17214	DYS276; 28296
	M188		DYS438; 13146	
O3a3c1a	M162	DYS257; 161		DYS627; 29943
O3a4	IMS-JST002611			DYS443; 38198 DYS288; 33216 DYS260; 36777
O3a4a	P103			DYS473; 47700
O3a5	M300		DYS380; 13730 DYS470; 20505	
P	P237		DYS688; 1107	DYS480; 34599 DYS532; 44574
	P207	DYS257; 946		DYS627; 30728
	P243			DYS627; 34817
	P226		DYS8; 1143	DYS568; 22698
	P244		DYS11; 7029	DYS434; 33492 DYS437; 34056
	P283		DYS445; 1781 DYS485; 5352	DYS215; 42199
	P284		DYS445; 20370	DYS485; 27503
	P235		DYS84; 20697	DYS444; 28117 DYS541; 41072
	P281		DYS80; 5415 DYS642; 8057	DYS577; 39511 DYS200; 40168
	P282	DYS587; 290	DYS476; 20144	DYS514; 38958
	P239		DYS606; 16702	DYS475; 31183
	P230			DYS643; 43982 DYS605; 28200
	P240		DYS148; 9004 DYS507; 12331 DYS389A; 13608 DYS467; 13731	DYS466; 37394
	P69		DYS438; 11595	
Q	M242	DYS66; 832	DYS275; 9199 DYS495; 7189 DYS440; 2583	DYS441; 36474
P36.2=Q1; p36.1=A2	P36.2=P36.1	DYS435; 34	DYS581; 8613 DYS439; 18969	DYS434; 29850 DYS437; 29286
Q1a1	N14		DYS275; 21267 DYS495; 19257 DYS440; 14651 DYS66; 12900	DYS441; 48542
Q1a3	M346		DYS504; 16284	DYS579; 44843 DYS251; 40489 DYS252; 27167 DYS536; 27967
Q1a3a	M3	DYS199; 180	DYS617; 14673	DYS484; 46797 DYS483; 38546
Q1a3a2	M194		DYS275; 5167 DYS495; 5157 DYS440; 1449 DYS66; 3200	DYS441; 32442
Q1a3a3	P106		DYS473; 7209	
	M199		DYS275; 21728 DYS495; 19718 DYS440; 15112 DYS66; 13361	DYS441; 49003
	p292=Q1a3a3; M320=T1		DYS275; 21384 DYS495; 19374 DYS440; 14768 DYS66; 13017	DYS441; 48659
Q1a4	P48	DYS435; 245	DYS581; 8824 DYS439; 18758	DYS434; 30061 DYS437; 29497
Q1b	M378		DYS275; 18124 DYS495; 16114 DYS440; 11508 DYS66; 9757	



haplogroup	SNP	<1kb	>1kb -25kb	>25kb - 50kb
R	P232			DYS573; 28720
	M306		DYS380; 12345 DYS470; 19120	
	P227		DYS494; 23406 DYS550; 3581	DYS585; 41355
	P285		DYS84; 48435 DYS444; 41015 DYS541; 28060	
	P224		DYS510; 14075 DYS588; 21315 DYS512; 21697 DYS511; 19004 DYS390; 10896 DYS721; 9961 DYS517; 7567 DYS723; 3903	DYS518; 34124 DYS717; 27398
	M207		DYS718; 21023 DYS276; 22103 DYS712; 23407	
R1	P294		DYS288; 9120 DYS260; 12681	
	P242		DYS651; 17285 DYS463; 3697	
	P238			DYS592; 47134 DYS520; 40580 DYS54; 26773
	P231		DYS269; 9763	
	P233		DYS415; 13522	
	P234			DYS414; 34948
	P236	DYS644; 567	DYS508; 11779 DYS501; 20242 DYS74; 20273	DYS243; 34889 DYS281; 27836 DYS28; 38657
	P286			DYS243; 31038 DYS281; 38091
	P225		DYS276; 13744 DYS275; 17041	DYS712; 31766 DYS441; 44316
	M173=P241		DYS495; 15031 DYS440; 10425 DYS66; 8674	
R1a	SRY10831.2		DYS376; 6152	DYS242; 34317
R1a1	M198		DYS275; 21369 DYS495; 19359 DYS440; 14753 DYS66; 13002	DYS441; 48644
R1a1c	M204		DYS276; 12885	DYS712; 32625
R1a1e	PK5		DYS653; 20132	
R1b	M343		DYS504; 15615	DYS579; 45512 DYS251; 41158 DYS252; 26498 DYS536; 27298
R1b1b	P297			DYS539; 34627
R1b1b2	M269		DYS380; 23561	DYS470; 30336
R1b1b2c	M153			DYS594; 49352 DYS548; 40559
R1b1b2d	RY2627=M16		DYS376; 5057	DYS242; 33222
R1b1b2e	USP9Y+3636=M222			DYS438; 35601
R1b1b2f	P66			DYS237B; 25599
R1b1b2g	U106=M405			DYS122; 32035 DYS568; 26604
R1b1c	M335		DYS275; 17012 DYS495; 15002 DYS440; 10396 DYS66; 8645	DYS441; 44287
R2	P249		DYS688; 1435	DYS480; 34271 DYS532; 44246
	P267	DYS587; 282	DYS476; 20136	DYS514; 38950

haplogroup	SNP	<1kb	>1kb -25kb	>25kb - 50kb
S1	M254		DYS275; 13380 DYS495; 11370 DYS440; 6764 DYS66; 5013	DYS441; 40655
S1a	P57	DYS581; 762	DYS434; 20475 DYS437; 19911 DYS435;9341	DYS11; 46938 DYS439; 28344
S1b	P61		DYS434; 15603 DYS437; 15039 DYS581; 5634 DYS435; 14213	DYS11; 42066 DYS439; 33216
S1c	P83			DYS265; 28532
S1d	M226		DYS276; 12639	DYS712; 32871
T	M272		DYS380; 24153 DYS275; 5123	DYS470; 30928 DYS441; 32398
	M193		DYS495; 3113 DYS440; 1493 DYS66; 3244	
	M184=USP9Y+3178			DYS438; 39852
T2	P77		DYS438; 11813	

Appendix-Figure 7-3: Nearest Y-STR marker of 347 Y-chromosomal phylogenetic SNPs in 191 different branches. Phylogenetic tree shows only branches (blue) when the distance to one or more Y-STR markers was smaller than 50 kb. SNPs and tree composition were taken from the last recent publication of the YCC [142]. Branches and SNPs in grey indicated SNPs with the same position but different polymorphisms and different branches. For the determination of Y-chromosomal SNP-STR pairs the SNP position and the up-stream border of each STR system were used (Appendix-Table 7-4). In total, 199 different Y-STR markers were found less than 50 kb near one SNP, from which 24 markers show less than 1 kb distance, 118 markers between 1kb and 25kb and 57 markers between 25 kb and 50 kb distance. The number behind a marker presents the calculated distance to the nearest phylogentic SNP in bp. Note that grey labeled SNPs and branches are not counted twice. Yellow colored DYS markers indicate markers, which appear more than one time in the near of a phylogenetic SNP.

7.5 List of publications

Publications of this work

In preparation: "Extraction of a combined haplotype from male mixed sample by a multiplex HSE approach"

J. Rothe, N.E. Jr Watkins, M. Nagy, New Prediction Model for Probe Specificity in an Allele-Specific Extension Reaction for Haplotype-Specific Extraction (HSE) of Y Chromosome Mixtures. PLoS One, 2012. **7**(9): p. e45955.

J. Rothe, L. Roewer, M. Nagy, Individual specific extraction of DNA from male mixtures--First evaluation studies. Forensic Science International: Genetics, 2011. **5**(2): p. 117-121.

J. Rothe, M. Nagy, Strategies for excluding false Y-chromosomal SNP entries from human genome databases. Electrophoresis, 2012. **33**(9-10): p. 1488-1491.

Presentation

„New probe design strategy for an improved specific extension reaction for HLA sequencing", **J. Rothe**, J. Dapprich, M. Nagy, 19 *Jahrestreffen der Deutsche Gesellschaft für Immungenetik*, September 22-24, 2011, Berlin, Germany

„*Entwicklung und Validierung von Sonden für die Trennung männlicher DNS-Mischungen mittels Haplotypspezifischer Extraktion*", **J. Rothe**, L. Roewer, M. Nagy, 31. *Spurenworkshop*, February 25-26, 2011, Hamburg, Germany

„Individual specific extraction of DNA from male DNA mixtures – first evaluation studies", **J. Rothe**, L. Roewer, M. Nagy, 7th international forensic Y-user workshop / 4th EMPOP meeting, April 22-24, 2010, Berlin, Germany

„*Zur Problematik Y-chromosomaler SNP-Datenbanken – erste eigene Rechercheergebnisse*", **J. Rothe**, M. Nagy, 30. *Spurenworkshop*, February 5-6, 2010, Zurich, Switzerland

Poster-Presentation

"Strategies for excluding false Y-chromosomal SNP entries from human genome databases", **J. Rothe**, M. Nagy, *Jahrestreffen der Arbeitsgemeinschaft für Gen-Diaknostik*, October 28-29, 2011, Potsdam, Germany

„Visual Omp™ software; An excellent tool for effective primer design in complex PCR reactions and allele-specific assays", **J. Rothe**, N.E. Jr. Watkins, M. Geppert, M. Nagy, 24th World-Congress of the international society of forensic genetics, August 28 – September 3, 2011, Vienna, Austria

8 ACKNOWLEDGMENTS

Ich möchte mich zunächst ganz herzlich beim kompletten Team der Forensischen Genetik Berlin für die vielfältige Unterstützung bedanken. Es hat mir durch ihre Hilfsbereitschaft geholfen Probleme jeglicher Art zu bewältigen.

Mein besonderer Dank gilt meiner Betreuerin Frau Dr. Nagy, die sich durch keinen Stress der Welt davon abhalten ließ, mir mit Rat und Tat zur Seite zu stehen. Ich danke Ihnen für die vielen konstruktiven Diskussionen und für die Möglichkeit eigene Ideen zu testen.

Ich möchte auch ganz herzlich meiner Kollegin Maria danken. Als besonders angenehme und hilfreich empfand ich immer unseren Erfahrungsaustausch, bei dem ich besonders bezüglich der Y-SNPs eine Menge von Dir lernen konnte. Weiterhin danke ich Petra Otremba für Ihre professionelle Unterstützung bei der HLA Sequenzierung.

Ein Danke geht auch an meinen internen Betreuer Herr Prof. Dr. Börner für sein Engagement und die Lösung organisatorischer Probleme.

Wie immer muss ich meinem besten Freund Niels danken, der mir diesmal bei der Eingabe hunderter von Allelflächen geholfen hat und auch sonst immer interessiert meinen Berichten zugehört hat.

Mein letzter Dank gilt meiner Familie sowie meinem liebevollen Freund, die mich stets bei all meinen Vorhaben unterstützen und motivieren.

Erklärung

§ 5 Abs. 2 der Promotionsord. der Mathematisch-Naturwisschenschaftlichen Fakultät I

Hiermit versichere ich, dass

- ich die vorliegende Dissertation selbständig und nur unter Verwendung der angegebenen Literatur angefertigt habe.
- die Arbeit an keiner anderen Einrichtung zur Begutachtung eingereicht wird bzw. wurde.
- ich mich bei keiner weiteren Einrichtung um einen Doktorgrad beworben habe.
- mir bei dem angestrebten Verfahren die zugrunde liegende Promotionsordnung bekannt ist.

Datum…….. Unterschrift……